- These slides/notes represent only part of the course, and were accompanied by face-to-face explanations on white-board and additional topics / learning materials.
- In preparation of these slides I have also benefited from various books and online material.
- Some of the slides contain animations which may not be visible in pdf version.
- Corrections, comments, feedback may be sent to https://www.linkedin.com/in/naveedrazzaqbutt/

ES 544
Random Processes

with

Dr. Naveed R. Butt

@

GIKI - FES

Introductions …

- Me
- You
- The Course

GIK Institute

BS in Engineering Sciences

[1998 - 2002]

Automation Engineer [2002 - 2004]

Riyadh Water Transportation System's SCADA upgrade project

KFUPM

MS Student/Staff [2004 - 2006]

- MS in Systems Engineering
- Thesis in nonlinear modelling & control
- Teaching (labs: DSP, Control)

LTH

[2006 - 2014]

- **Positions: PhD Student/Staff, Postdoc, Research Associate**
- **PhD in Engineering (focus: Statistical Modelling & SP)**
- **Teaching + Research**

Ericsson Research

[2014 - 2018]

- **Senior Researcher**
- **Research + Patenting**
- **Next Generation WiFi & 5G**

Jouf University [2018…]
Assistant Professor
College of Engineering

- **Badminton, Bowling**
- **Weekend dinners**
- **Reading, Writing (poetry, short stories)**

Roles

Teacher

Researcher

As a teacher

Full Courses

- Stochastic Processes
- Statistical DSP & Modelling
- Probabilistic Methods in Engineering
- Wave Propagation & Antennas
- Principles of Communications
- Digital Communications
- Satellite Communications
- Circuit Analysis II

Labs & Tutorials

- Time Series Analysis
- Signal Theory
- Advanced Control
- Modern Control Systems
- Digital Design

Supervision

Supervised and collaborated in various grad and postgrad theses.

As a researcher

Statistical Modelling of Signals & Systems

Spectroscopy

- Nuclear Quadrupole Resonance (NQR) signal detection
- Raman Signal Classification

Spectrum Estimation

- Missing Samples Cases
- Poly-spectra
- Coherence Spectra

Beamforming

- Radar & Sonar
- Pitch Estimation

Communications

- WiFi & 5G
- Antenna Arrays

Control

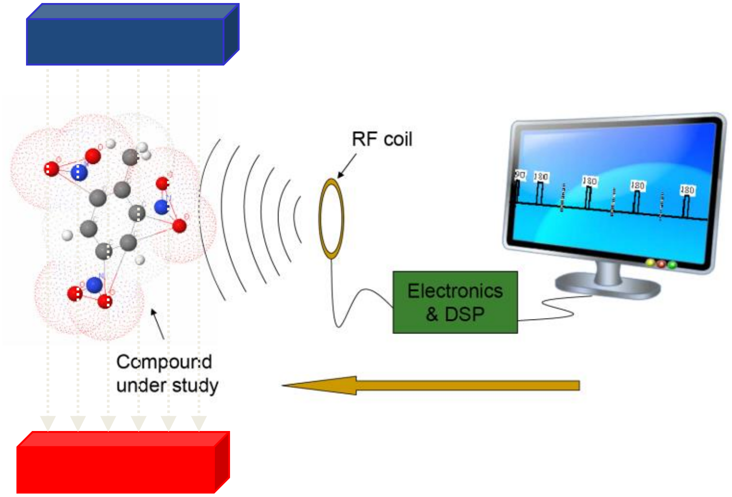
- Nonlinear Plant Modelling & Control

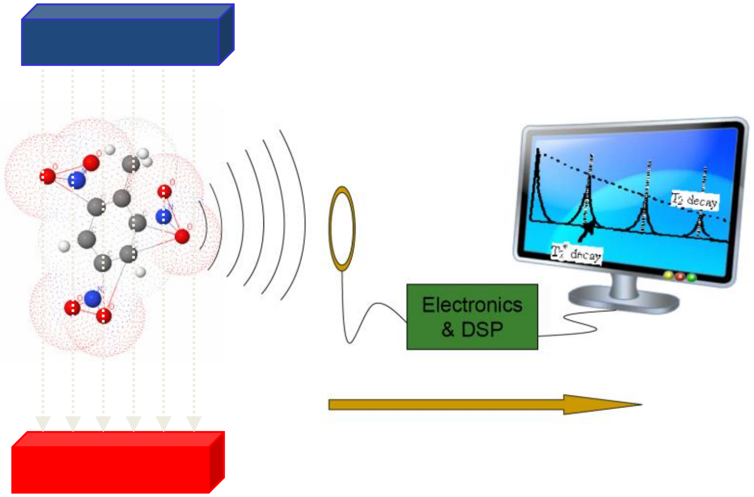
One of my research projects…

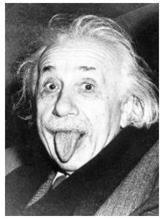
CONPHIRMER Project

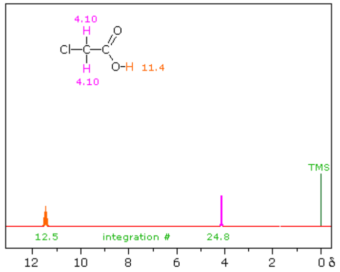
Question: how to quickly tell whether a medicine is fake?

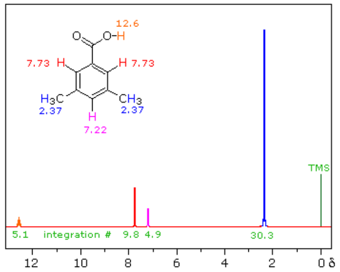
NMR vs. NQR


RF coil
Electronics & DSP
Compound under study

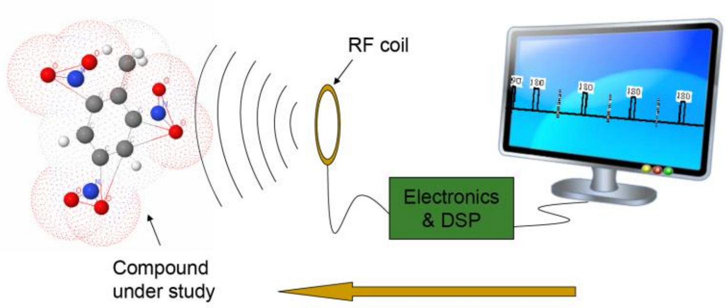

Electronics & DSP

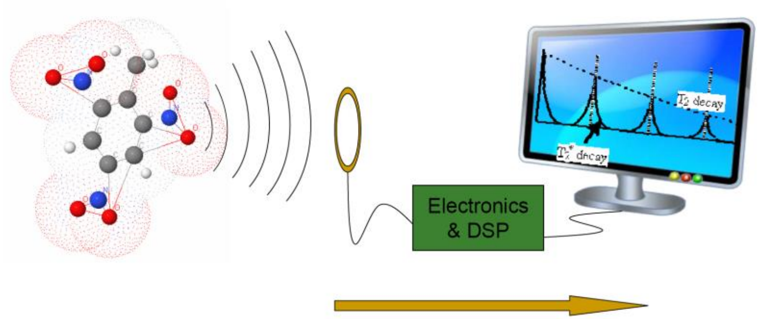


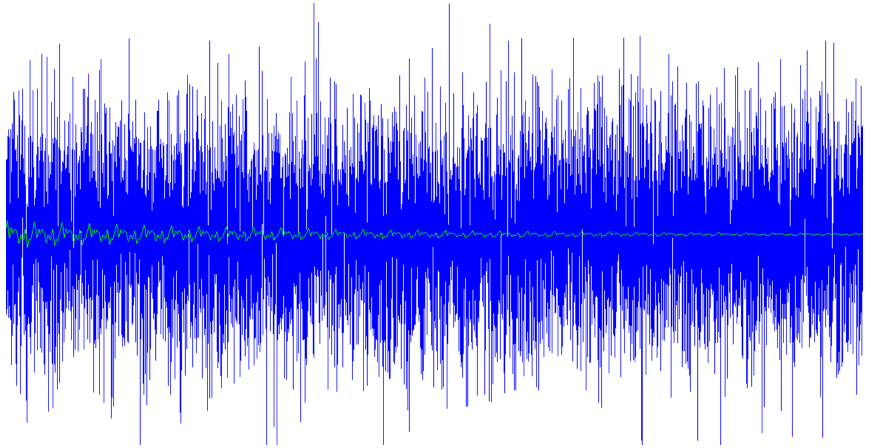

TMS
integration #
12 10 8 6 4 2 0 δ


TMS
integration #
12 10 8 6 4 2 0 δ

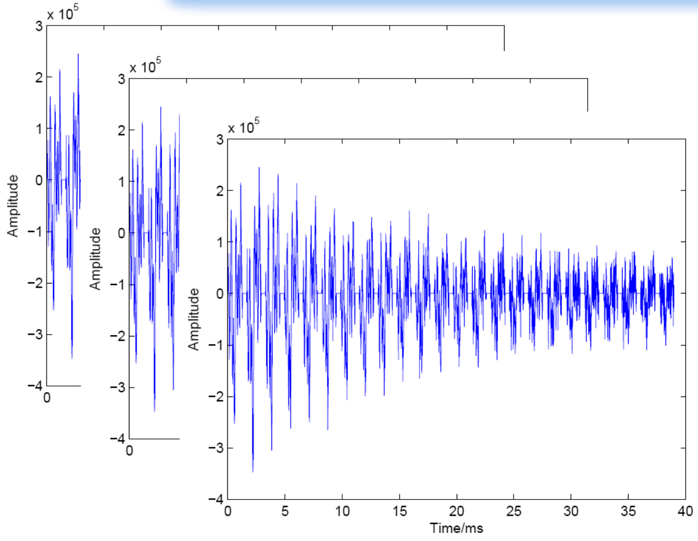
NQR Signal

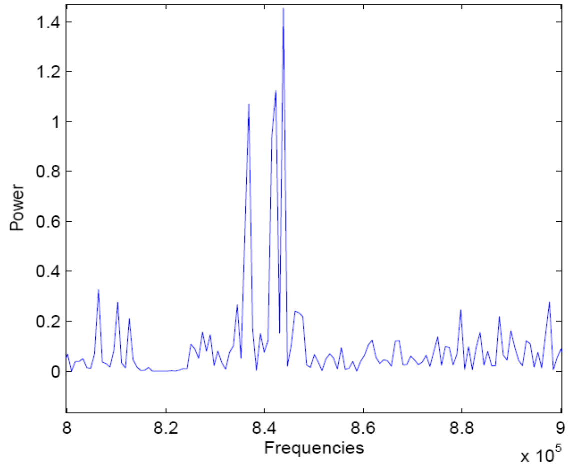

RF coil
Electronics
& DSP
Compound
under study


Electronics
& DSP



NQR Signal Modeling & Detection

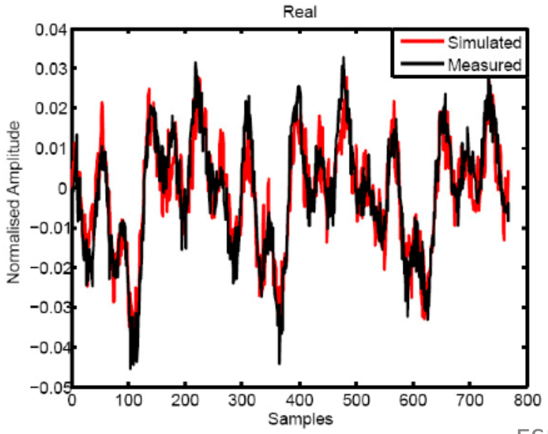




$$y_m^{(p)}(t) = \sum_{k=1}^{d^{(p)}} \alpha_k^{(p)} e^{-(t+m\mu)\eta_k^{(p)}} e^{-\beta_k^{(p)}|t-t_{sp}|+i\omega_k^{(p)}(T)t}$$

$$\omega_k^{(p)}(T) = a_k^{(p)} - b_k^{(p)} T$$

$$\alpha_k^{(p)} = \rho\kappa_k^{(p)} \qquad \left\|\boldsymbol{\kappa} - \boldsymbol{\kappa_a}\right\|_2^2 \leq \epsilon$$



Introductions …

- Me
- You
- The Course

The Course …

- Why study random processes?
- Where can I apply what I learn?
- How is the course conducted and assessed?

The Course …

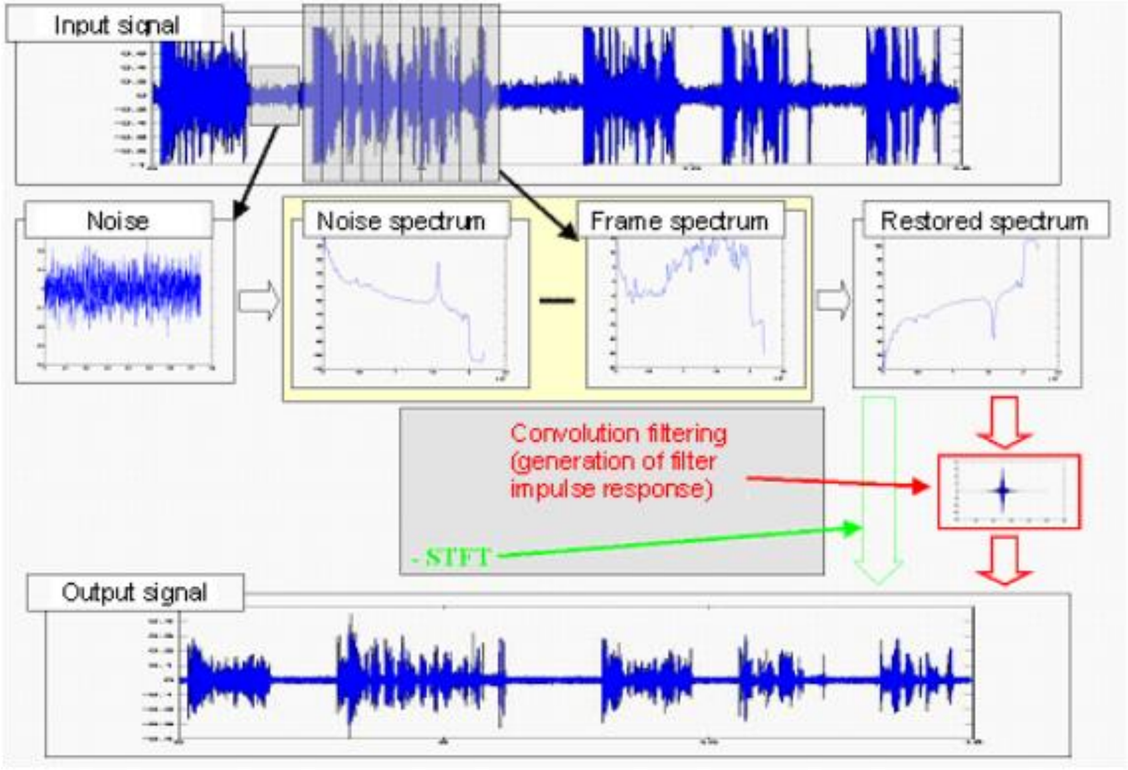
- Why study random processes?
- Where can I apply what I learn?
- How is the course conducted and assessed?

The Course …

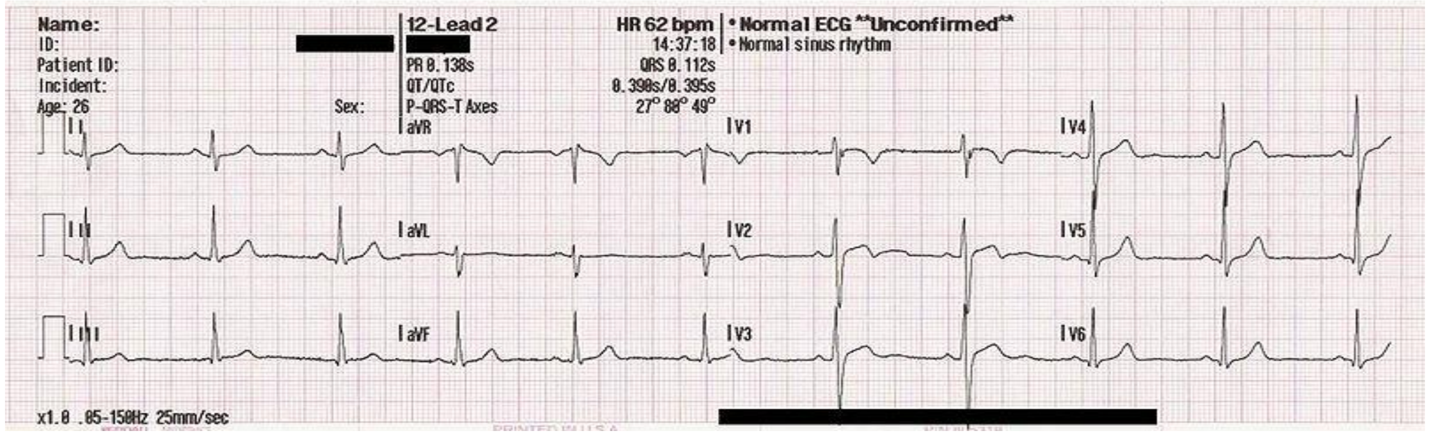
- Why study random processes?
- Where can I apply what I learn?
- How is the course conducted and assessed?

Basically, wherever you would like to (or would benefit from) including **uncertainty or variations** in your **mathematical model**.

Communications



Bio-Medicine

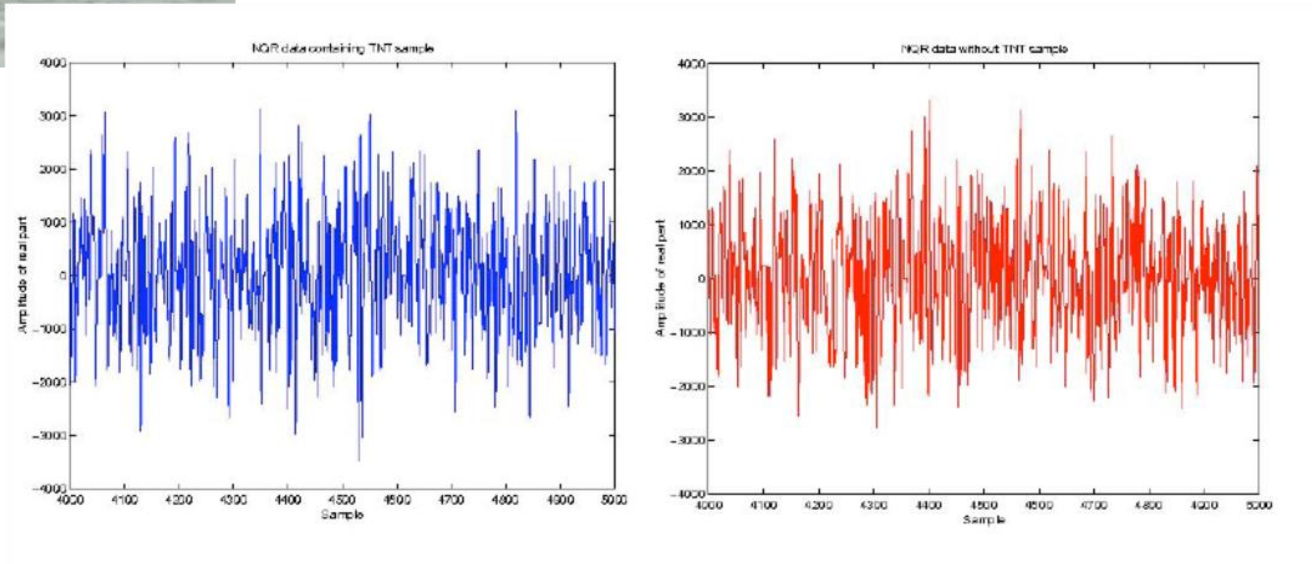


Finance

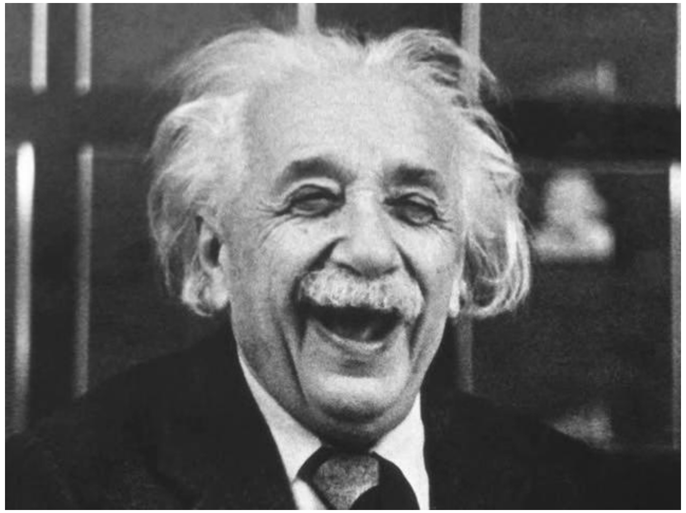
...WITH OUR NEW FINANCIAL REGULATIONS IT'S NOW SAFE TO INVEST YOUR MONEY...
MY WHAT?
HELP!

XLF (Financials Select Sector SPDR) AMEX
©StockCharts.com
10-Aug-2007 Open 32.19 High 35.19 Low 31.98 Close 33.35 Volume 519.8M Chg +1.29 (+4.02%)
XLF (Weekly) 33.35
MA(50) 36.00
MA(200) 30.78
MA(20) 36.27
Volume 519,772,480
600M
500M
400M
300M
200M
100M
38
37
36
35
34
33
32
31
30
29
28
27
26
S O N D 05 F M A M J J A S O N D 06 F M A M J J A S O N D 07 F M A M J J A

Chemistry


NQR data containing TNT sample
Amplitude of real part
Sample
NQR data without TNT sample
Amplitude of real part
Sample

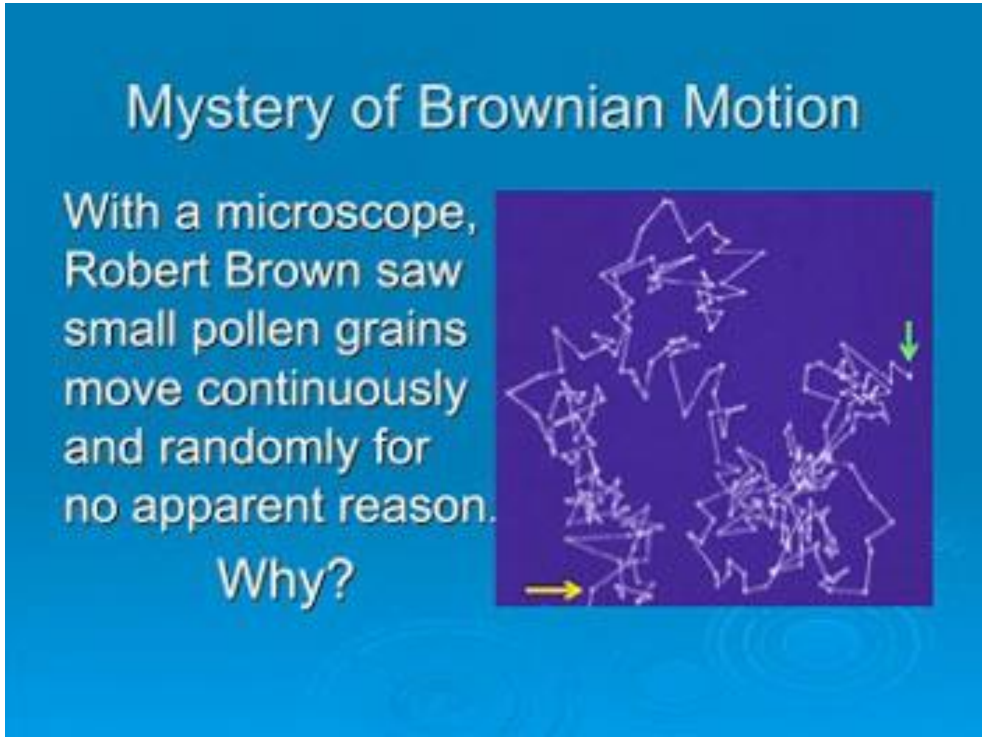
Or, just use it to prove existence of atoms…

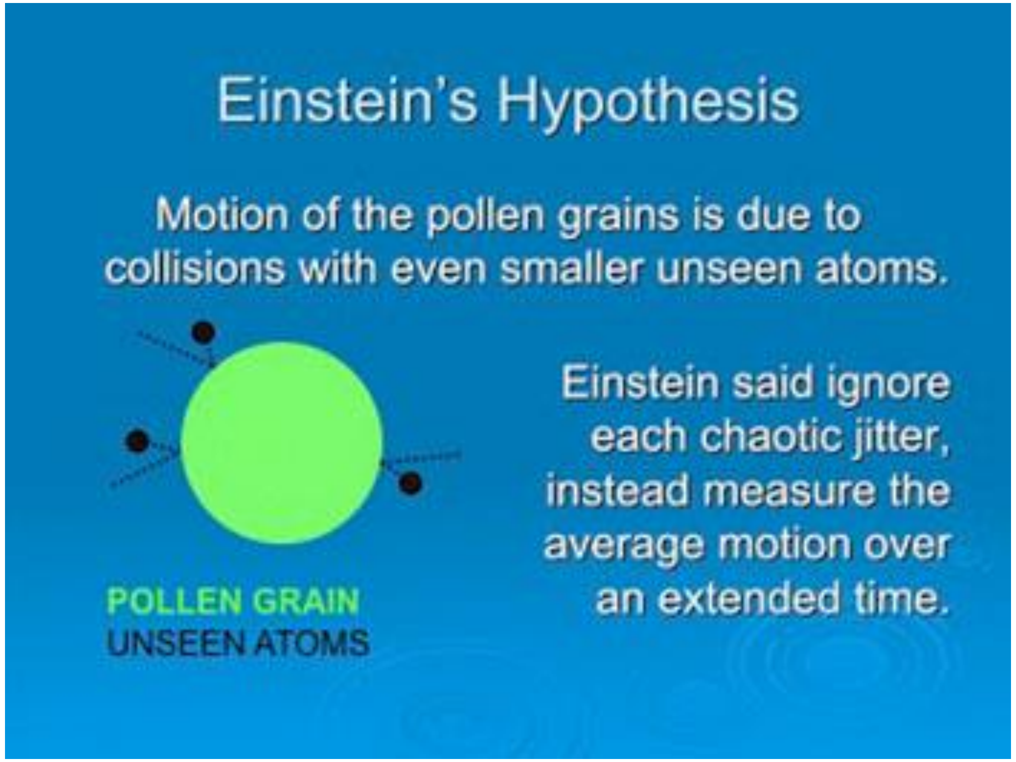


Einstein's 1905 papers

1. *On a Heuristic Point of View on the Creation and Conversion of Light* (Photo-Electric Effect)
   http://lorentz.phl.jhu.edu/AnnusMirabilis/AeReserveArticles/ein
2. *On the Electrodynamics of Moving Bodies* (Theory of Special Relativity)
   http://www.fourmilab.ch/etexts/einstein/specrel/www/
3. ***Investigation on the Theory of the Brownian Movement***
   http://lorentz.phl.jhu.edu/AnnusMirabilis/AeReserveArticles/ein
   pdf

Or, just use it to prove existence of atoms…





The Course …

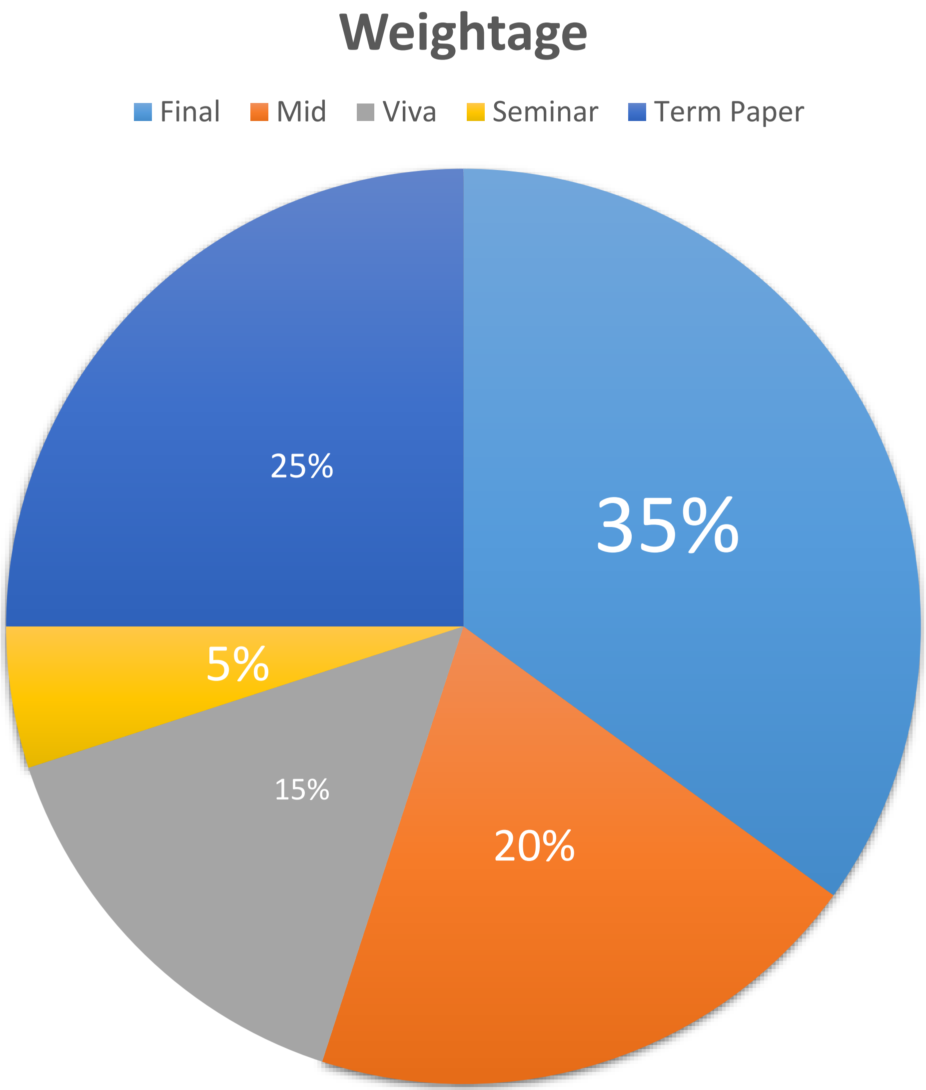
- Why study random processes?
- Where can I apply what I learn?
- How is the course conducted and assessed?

Important Business!!

- 80% attendance is mandatory!
- Textbooks
  - **Lindgren**, G., Rootzen, H., Sandsten, M., *Stationary Stochastic Processes for Scientists and Engineers*, CRC Press, 2013
  - **Kobayashi**, H., Mark, B., & Turin, W., *Probability, Random Processes, and Statistical Analysis*, Cambridge University Press, 2011.
- Contact
  - naveed.butt@giki.edu.pk
  - office: FES G-6

Learning Plan

- **Lectures**
  - Help discover and grasp new concepts
- **Viva (Assignment-based)**
  - Help prepare/revise each week's concepts
  - Keep you from lagging behind in course
- **Seminar**
  - Prepares for presentations and teaching
- **Term Paper**
  - Helps practice independent research/review work
  - Helps appreciate wider aspects of course content
- **Exams (Mid, Final)**
  - Help prepare entire course material

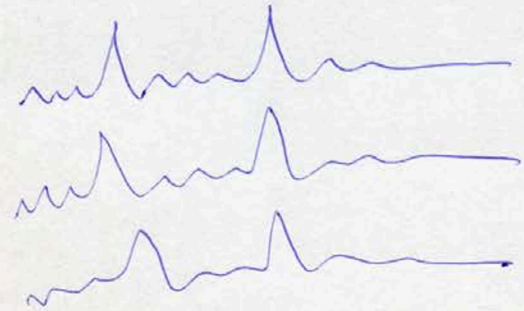


Questions?? Thoughts??

Random processes Lec 1 (intro)

"I hate this course".

— Introductions
  — Me (Dr. Naveed R. Butt)
  — You
  — The course.



— why study random processes?

suppose — record same utterance thrice
— will the corresponding samples be exactly the same?
— There will be random variations between the recordings.
— But our ears can still tell alot (— same person — same sentence)
— Random processes is, in part, study of such variations/uncertainties to possibly extract underlying deterministic information.

— First, we would be interested in seeing how to represent the random variations and their distributions.

→ That's chapter 1 : Random Processes (Basics, Distributions etc.)

— We would be interested in extracting some non-random (deterministic) info from the process, e.g. average value, degree of variation, co-dependance and in seeing if those are fixed or time-varying.

→ That's chapter 2: Moments & Stationarity

— For ease of analysis and comparison, we would be interested in studying properties of commonly occuring random processes.

→ chapters 3 $ 5 : Common Random processes- (Poisson, Gaussian, Wiener, Markov)
↳ 2nd TextBook

— We might also want to see what mathematical functions can represent random processes efficiently (helps in compression, interpolation, prediction, etc...)

→ Chapter 7 : ARMA Models.

— We could be interested in developing mathematical operations that can controllably alter ~~or~~ the stochastic process.

→ chapters 6 $ 8 (— Filter Basics
— Integrators, Differentiators
— Optimal Filters)

— Quite often, it helps to see phenomena from a different angle. E.g., most natural phenomena have underlying vibrational motions. It could be useful to see a phenomenon in terms of its vibrational frequencies.

→ Chapter 4 $ 9 (— spectral representation
— spectral Estimation)

① what is probability?

- Probability is the lack of knowledge!
- you know your height, but what's the height of the next person entering the room?

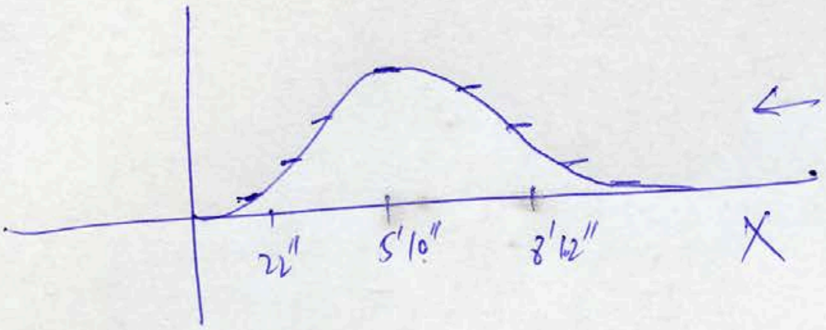
(1.1) why do we sometimes lack knowledge?

- Future (die I haven't rolled yet)
- Too hard to collect information (drone following everyone)
- Variations (e.g. — handwritings — Different photos of same person)
- Quantum randomness (some behaviors at quantum level are random by their very nature, and not because of our models or instrument limitation)

current concenses →

② what is statistics (or statistical Analysis)?

- It is a tool that helps us make good use of whatever knowledge we have!
- e.g. our lack of knowledge may not be absolute
- we do know something about height of next person to enter the room.



← a graph of relative likelihoods of heights..

(3) How do we assign probabilities?

— Repeated experimentation / observation
(i.e via some collected statistics)

— Belief.

L2

(9)

Four steps

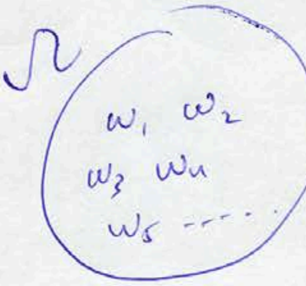
(9.1) Define your experiment clearly.

→ e.g. rolling a die

(9.2) Define the sample space (set of all distinct possibilities)

$\rightarrow \Omega = \{ \overset{\omega_1}{1} \; \overset{\omega_2}{2} \; \overset{\omega_3}{3} \; \overset{\omega_4}{4} \; \overset{\omega_5}{5} \; \overset{\omega_6}{6} \}$

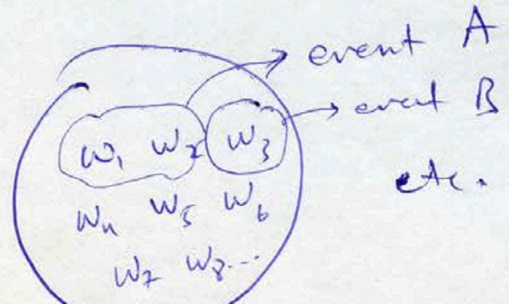
(e.g for die rolling).



(9.3) Define events

→ Here events = subsets of Ω including ϕ and Ω itself.

$A = \{ \omega : \omega \text{ satisfies some condition} \}$



probability measure (9.4) Assign probabilities (numeric values, real numbers) to events such that the assignment scheme makes sense (i.e. satisfies some axioms)

— Axiom 1: Assigned probabilities should not be negative!

$\Rightarrow P(A) \geq 0$ for all events A.

Axiom 2 : Probability of the sure event must be 1.

$$\Rightarrow P(\Omega) = 1$$

(also includes that : sample space must cover all distinct possibilities)

Axiom 3 : Probabilities assigned to mutually exclusive events (events that cannot occur at the same time) must make sense.

$\Rightarrow$ For mutually exclusive A and B, we must have

$$P(A \cup B) = P(A) + P(B)$$

(no overlapping terms etc.)

(4) What is joint probability?

$\rightarrow$ Sometimes we are interested in finding probability of two events occuring at the same time. This is called joint probability.

$$P(A \text{ and } B) = P(A \cap B)$$

$A = \{\text{Ali is in lecture}\}$ $B = \{\text{Ali is sleeping}\}$

$P(A \text{ and } B)$ = Probability that Ali is in lecture and sleeping.

(5) what is conditional probability?

→ sometimes knowledge of one random event can help us assign probability to another random event.

→ For these "conditional probability" comes in handy.

ex1 → A = {Ali is in lecture at 8 am}

B = {Ali is sleeping at 8 am}

Suppose I tell you that A has occurred (that is Ali is in lecture at 8 am), now what is the probability the Ali is sleeping?

$P(B \text{ given } A) = P(B|A)$

ex2 → Suppose I write an ~~number~~ integer from 1 to 4 on a piece of paper but do not tell you what I wrote. (except that it is an integer from 1 - 4)

→ what is the probability that I wrote 3?

$A = \{\text{wrote three}\}$ $P(A) = 1/4$

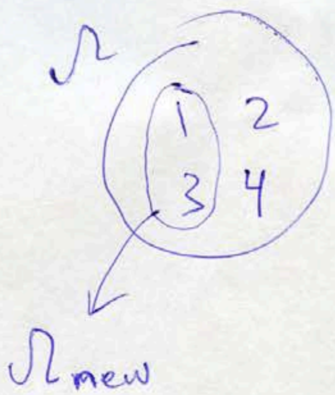
~~what is the~~

→ Suppose I tell you that I wrote an odd number

$B = \{\text{wrote odd number}\}$ ✓

now what is the probability that I wrote 3?

$$P(A \text{ given } B) = \frac{1}{2}$$



→ Addition of useful information has decreased the lack of knowledge (i.e ~~dec~~ increased the probability).

→ we write this as

$$P(A \text{ given } B) \triangleq P(A|B)$$

→ and link it to joint probability as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{given } P(B) \neq 0$$

Interesting → Note that if B cannot occur (i.e if $P(B)=0$) then the question "Probability A given that B has occurred" is senseless to begin with!

⑥ What is statistical independence?

→ when knowledge of one event does not change probability of another event, we say the two are statistically independent.

→ i.e. when knowledge of one event adds no useful information towards possibility of another event.

e.g. → Say I tell you that I wrote a positive number (in addition to already told: it's b/w 1 and 4).

$C = \{\text{wrote a positive number}\}$

Now what is the probability that I wrote 3?

$A = \{\text{wrote } 3\}$

$$P(A|C) = P(A)$$

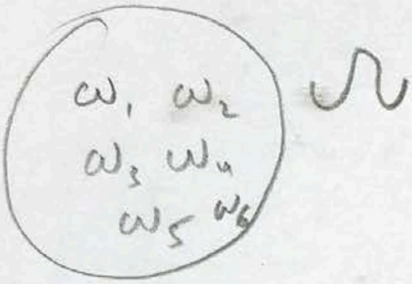
→ Two ways of testing independence

$P(A|B) = P(A)$ Test 1 (i.e. if conditional prob. same as unconditional)

$P(A \cap B) = P(A)\,P(B)$ Test 2 (i.e. if joint probability is separable into simple product

follows from $P(A|B) = \frac{P(A \cap B)}{P(B)}$ and test 1

— Some practice, Definitions, and Discussions



ω's are distinct "simple" events.

Assumed from here on!

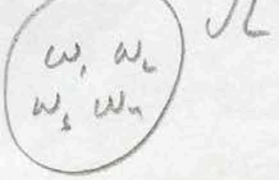
$P(\omega_1) = ?$

— we don't know from here, unless we say that the events are all equally likely!

— i.e. $P(\omega_1) = P(\omega_2) = \cdots = P(\omega_6)$ ← i.e have same probability of occurring.

— Then what is $P(\omega_1) = ?$ (Ans: $P(\omega_1) = 1/6$)

— Let's say



ω's are distinct simple events.

→ and I give you $P(\omega_1) = \frac{1}{5}$, $P(\omega_2) = \frac{1}{5}$, $P(\omega_3) = \frac{1}{5}$

— Find $P(\omega_4)$

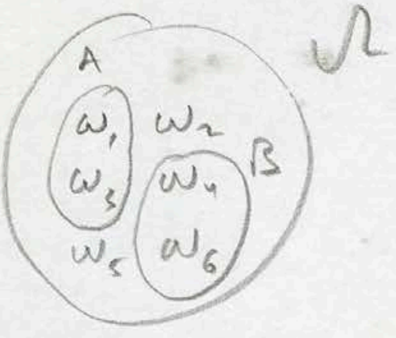
— Are these equally likely?

Ans: since $P(\omega_1) + P(\omega_2) + P(\omega_3) + P(\omega_4) = P(\Omega) = 1$

$$\Rightarrow P(\omega_4) = 1 - \frac{1}{5} - \frac{1}{5} - \frac{1}{5} = \frac{5}{5} - \frac{3}{5} = \frac{2}{5}$$

— $\omega_1, \omega_2, \omega_3$ are equally likely, ω_4 is not

— Another example ("Mutually Exclusive")



note: A and B are compound events (i.e. not 'simple')

— can A and B occur at the same time?

— NO
— Such events, as we mentioned briefly before, are <u>Mutually Exclusive</u>

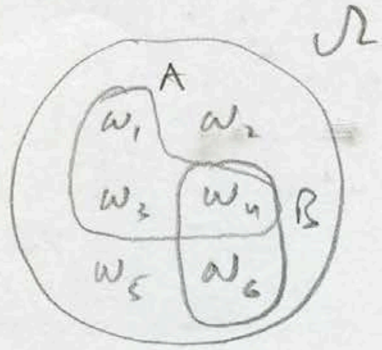
$P(A \text{ and } B) = P(A \cap B) = 0$ — cannot occur at the same time.

— note: 'Simple' events are always mutually exclusive (by definition)

— e.g. roll a die, define $A = \{\text{even number shows}\}$
i.e. $A = \{2, 4, \text{ or } 6 \text{ shows}\}$
$B = \{\text{odd number shows}\}$
i.e. $B = \{1, 3, \text{ or } 5 \text{ shows}\}$

— $P(A \cap B) = 0$ (number showing cannot be both even and odd
→ Mathematically impossible.)

— But what if



— Now A and B are not mutually exclusive.

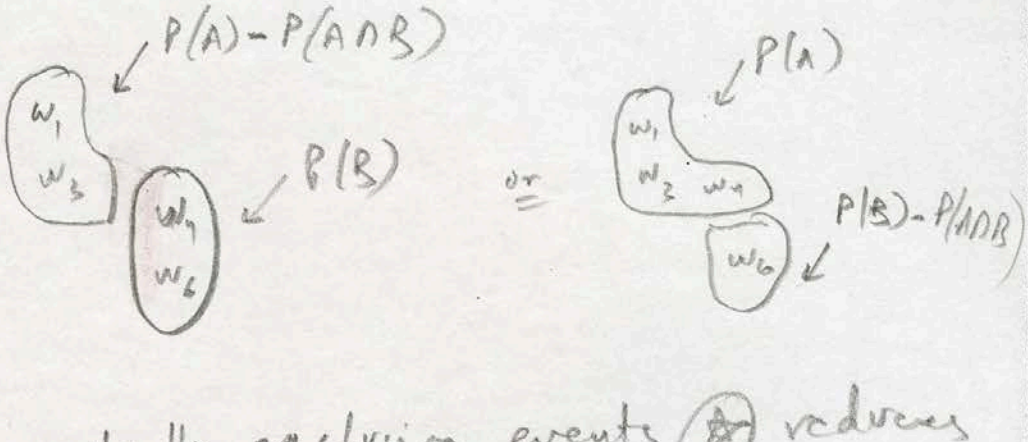
— $P(A \text{ and } B) = P(A \cap B) \neq 0$

— $P(A \cap B) = ?$

useful [$P(A \cup B) = P(A) + P(B) - \underbrace{P(A \cap B)}$ (☆)

↙ removes the overlapping part that was counted twice.

— It's something like



— note that for mutually exclusive events (☆) reduces to $P(A \cup B) = P(A) + P(B)$ (b/c there is no overlapping part to remove)

→ Recall: this was third axiom of probability

— Yet Another example (and "Bayes' Theorem")

— Assume that a certain system will collapse if two of its distinct components α and β both fail. If probability that α fails is 0.01, probability that β fails is 0.005, and probability that β fails if α has failed is 0.015.

Q

— Always define events: let $A = \{\alpha \text{ fails}\}$, $B = \{\beta \text{ fails}\}$

— Given i.e. $P(A) = 0.01$, $P(B) = 0.005$, $P(B|A) = 0.015$

$$P(A \cap B) = P(B|A)\,P(A)$$

Q. → What is the probability that A will fail if B has failed?

— i.e. $P(A|B) = ?$ → $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$ ← Now use

Actually = "Bayes' theorem" $\left[\; P(A|B) = \dfrac{P(B|A)\,P(A)}{P(B)} \;\right] = \dfrac{0.015 \times 0.01}{0.005} = 0.03$

→ Bayes' Theorem is very useful in finding conditional probabilities

→ It also is philosophical foundation of "Bayesian statistics"

→ used heavily in "learning systems"

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

$P(A)$ ← a priori probability of A

$P(A|B)$ ↓ posterior (or "updated" probability of A)

— And another example ("Total probability")

— Suppose your team is to play a PSL match against one of three teams: Peshawar Zalmi, Quetta Gladiators, or Lahore Qalandars. Your chances of winning from Zalmi are 0.05, from Gladiators are 0.04, and from Qalandars are 0.1.

Q: — What is your probability of winning a match if your opponent is chosen randomly with equal probability?

— Let's call our team Topi Drama and define events $W = \{\text{Drama wins}\}$

$Z = \{\text{Zalmi chosen}\}$, $G = \{\text{Gla— chosen}\}$, $Q = \{\text{Qalandar chosen}\}$

mutually exclusive.

— we are given

$$P(W|Z) = 0.05, \; P(W|G) = 0.04, \; P(W|Q) = 0.1$$

— and we want to find $P(D)$.

— Logically $P(W) = \begin{cases} \text{You play Z and win} \\ \text{or} \\ \text{You play G and win} \\ \text{or} \\ \text{You play} \end{cases}$ — also mutually exclusive

$$\Rightarrow P(W) = P(Z \cap W) + P(G \cap W) + P(Q \cap W)$$

or

$$P(W) = P(W|Z)P(Z) + P(W|G)P(G) + P(W|Q)P(Q)$$

$$= \frac{0.05 + 0.04 + 0.1}{3} = \frac{0.19}{3}$$

(note $P(Z) = P(G) = P(Q) = \frac{1}{3}$)

→ here we actually used something called "Total Probability" which says:

→ If A_k's are mutually exclusive events spanning Ω (i.e $\bigcup_{\forall k} A_k = \Omega$) then

$$P(B) = \sum_{\forall k} P(B|A_k)\, P(A_k)$$

Lecture 3

1. Embracing Uncertainty / variation : Prefix "Random"

Deterministic	→ Random
Event	Event
Variable	Variable
Process	Process

- event : moon passing between earth and sun (deterministic)
  - : Earthquake (random)
  - : Next student who enters being taller than me (random)
  - : Rolled die shows even number
  - : Next student entering is from Swabi region
- random event : An outcome of an uncertain happening/experiment

- Variable : Placeholder that helps give relations or hold data in algebraic form.
  - : $a = \pi r^2$
  - : X = height of the tallest person alive (as per records)
- Random variable : X = height of the next student entering the room
  - : Let's say radius of circle is decided based on outcome of a random experiment (e.g. rolling a die)

: Then r becomes a random variable (so does area)

: To emphasize dependance on a random event, we may rewrite it as

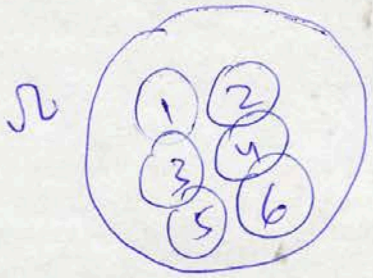
$r(\omega)$ ω = An event based on die rolling.

with associated $(\Omega, \sigma\text{-field}, P)$ ⊛

: There can be many ways of defining ω and assigning $r(\omega)$ values based on ω (mapping)

e.g.

Mapping 1



ω = Number showing on the die

$$r(\omega) = \omega \Rightarrow r(\omega) = \begin{cases} 1 & \omega = 1 \\ 2 & \omega = 2 \\ 3 & \omega = 3 \\ 4 & \omega = 4 \\ 5 & \omega = 5 \\ 6 & \omega = 6 \end{cases}$$

with ⊛

Mapping 2

ω = Number showing on the die

$$r(\omega) = \omega^2 \Rightarrow r(\omega) = \begin{cases} 1 & \omega = 1 \\ 4 & \omega = 2 \\ 9 & \omega = 3 \\ 16 & \omega = 4 \\ 25 & \omega = 5 \\ 36 & \omega = 6 \end{cases}$$

with ⊛

Mapping 3

ω = Number showing on the die with ⊛

$A = \{\omega : \omega \text{ even}\} = \{2, 4, 6\}$

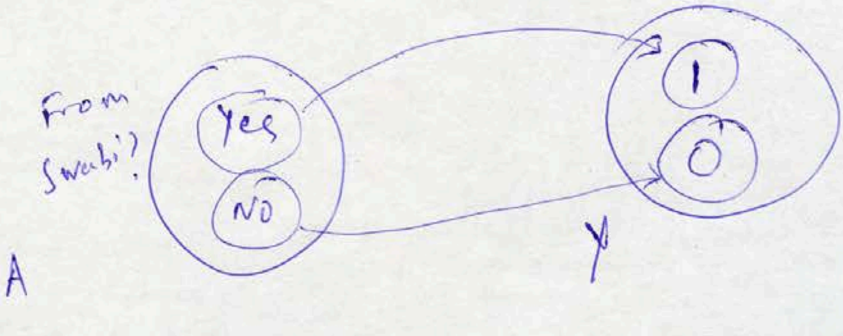
$B = \{\omega : \omega \text{ odd}\} = \{1, 3, 5\}$

$$r(\omega) = \begin{cases} 0 & \omega \in A \\ 1 & \omega \in B \end{cases}$$

Defn: → So, a random variable is a variable that takes a numeric value based on a random event.

more later:
- It can be discrete : as r(ω) is in our three mappings.
- or, it can be continuous : as X (height) is in our ~~mapping~~ example.
- But it cannot have a non-numeric value!

Random Event vs Random Variable →
- Random variable is based on a random event.
- Random event can have a non-numeric value, but random variable cannot.
  - e.g. Is next student entering the room from Swabi region?
- Non-numeric valued random event can still be converted into random variable by a mapping



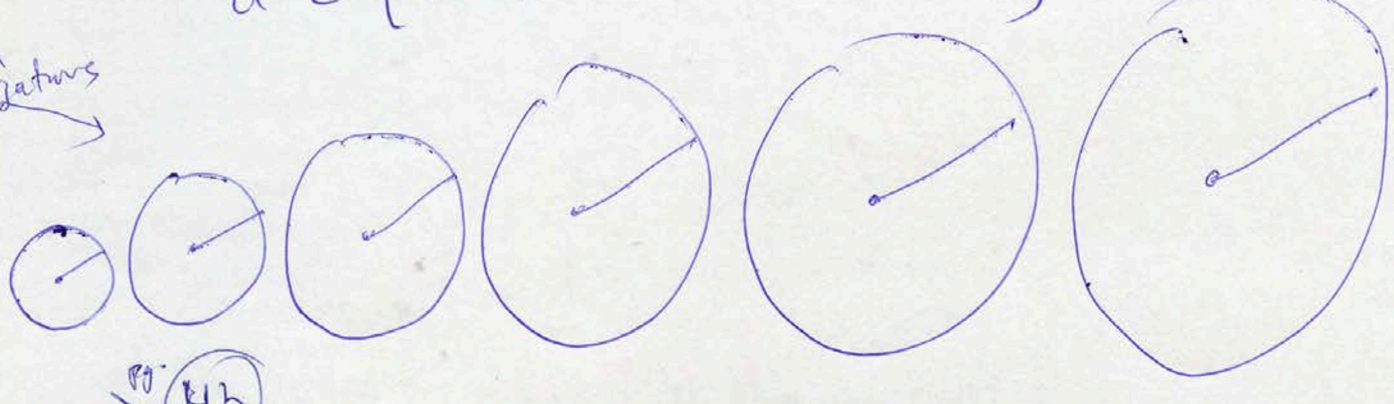
→ Realization : Once the outcome is available, the value of the random variable becomes fixed and we call it a realization

— Possible realisations of the circle in mapping 1

$$r \in [1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6]$$

$$a \in [\pi \quad 4\pi \quad 9\pi \quad 16\pi \quad 25\pi \quad 36\pi]$$

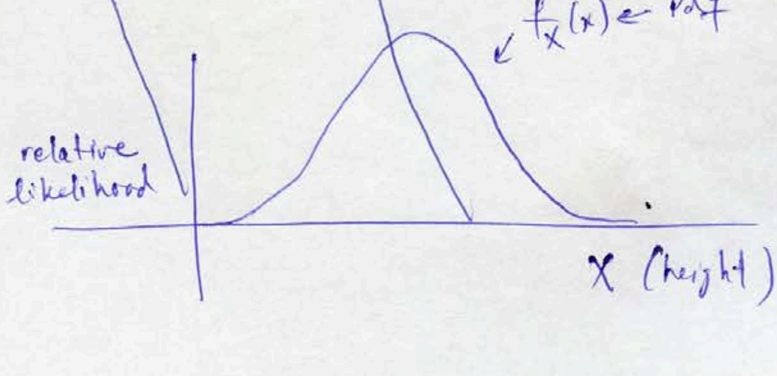
Area realizations →



→ pg. 49b

(2) Making Sense of an uncertain world!

- Probability is the lack of knowledge / uncertainty
  - e.g. height of next person walking in
- Statistics is a tool that helps make good use of whatever knowledge we have about the relative likelihoods of the events.
  - This knowledge is collected in density or mass functions.

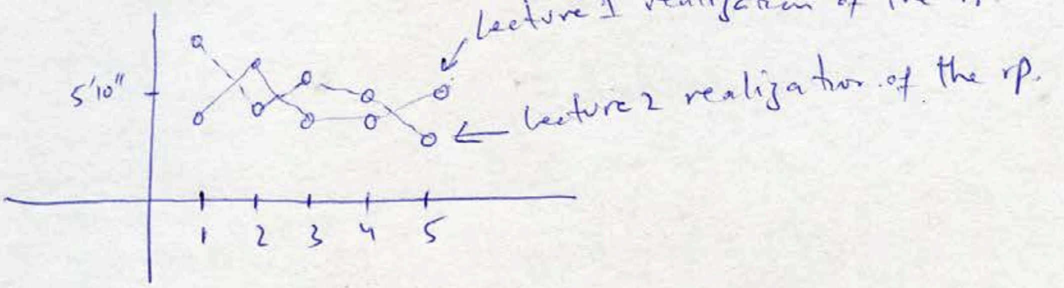


height of the next student entering the room.

— Random process: a collection of random variables

X = height of next student entering (rv)

(rp) ⇒ $[X_1\ X_2\ X_3\ X_4\ X_5]$ heights of next five students entering



— more generally, we may write

$$X(\omega, t) \qquad t \in T$$

$t \in T$: Discrete vs. continuous] more on this later.

ω: Discrete vs. continuous

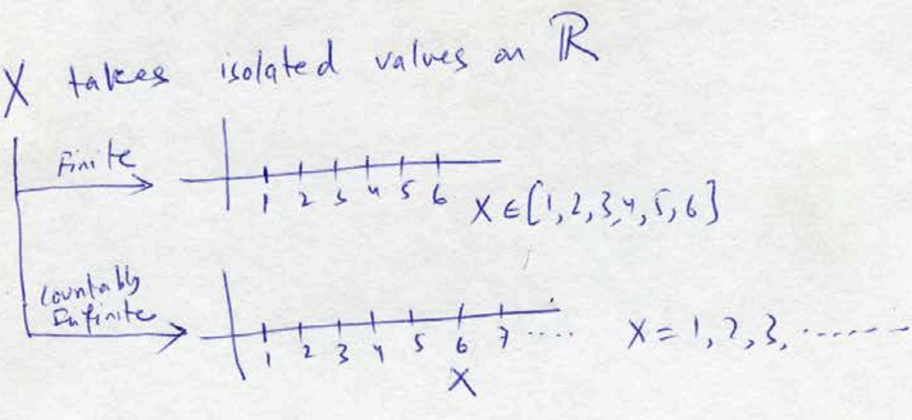
e.g. (How many siblings next student has? < Discrete in t, Discrete in value)

(2) Assigning Probabilities ("And Collecting") ($ related Functions)

Number of values X can take are (cases)

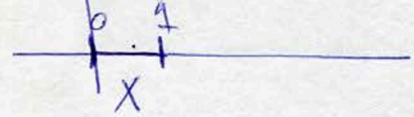
— Finite] Discrete random variable
— Countably Infinite] Discrete random variable
— Uncountably Infinite] e.g. continuous random variable

Discrete r.v. → X takes isolated values on $\mathbb{R}$



Continuous r.v. → X takes a continuum of values on $\mathbb{R}$

e.g → $X \in \mathbb{R}$

e.g → $X \in [0, 1]$ 

→ PD & PMF for Discrete rv

— for a discrete rv, we may collect the probability in a Prob. distribution (PD) as
with values $[x_1, x_2, x_3 \dots]$

→ $P_X(x_i) \triangleq P(X = x_i)$ $x_i \in [x_1, x_2, x_3 \dots]$ ~~i = 1, 2, 3, ...~~ (called PD)

— Another way is to define the Prob. Mass function (PMF) so that $x \in \mathbb{R}$ and $P_X(x) \in [0, 1]$

→ $P_X(x) \triangleq P(X = x)$ $x \in \mathbb{R}$

→ with the implicit assumption that $P_X(x) = 0$ whenever $x \notin [x_1, x_2, x_3 \dots]$

note 1 — of course PD and PMF must satisfy axioms of Prob. (so these have same properties as $P(\cdot)$)

note 2 — A countably infinite ranged ~~discrete~~ rv can still be assigned probabilities as long as they ~~are not uniform~~ satisfy the axioms,

→ e.g ~~for~~ let's say rv X can have values $k = 1, 2, 3 \dots$

→ then distribution $P[X = k] = \left(\frac{1}{2}\right)^k$ $k = 1, 2, 3, \dots$ can be ok

e.g $\sum_{k=1,2,3\dots} \left(\frac{1}{2}\right)^k = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1$ (recall $\sum_{k=1}^{\infty} x^k = \frac{x}{1-x}$ $|x| < 1$)

→ but uniform distribution will not be allowed by Kolmogorov (interesting: see "De Finetti Lottery")

e.g. if you set $P[X=k] = \alpha \quad k = 1, 2, 3, \ldots$

then $\sum_{k=1}^{\infty} \alpha = \begin{cases} \infty & \text{if } \alpha > 0 \\ 0 & \text{if } \alpha = 0 \end{cases}$

— We often find it useful to discuss whether an rv lives in a certain range

— Prob. next student entering is shorter than me?

— This is done with the help of CDF

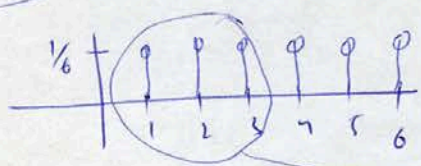
$$F_X(x) \triangleq P[X \leq x]$$

— For D.rv this simply translates to

$$F_X(x) = \sum_{x_i \leq x} P_X(x_i)$$

— e.g. for die rolling we have $\boxed{x_i = i}$ $i = 1, 2, 3, 4, 5, 6$

and $P_X(x_i) = \frac{1}{6} \quad \forall x_i$



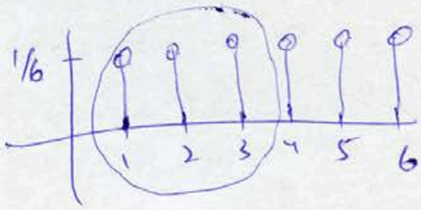
Then $F_X(3) = \sum_{x_i \leq 3} P_X(x_i) = P_X(1) + P_X(2) + P_X(3) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6}$

$F_X(4) = ?$

— so The CDF

$$P(X \leq a) \triangleq F_X(a)$$

$\frac{1}{6} + \frac{1}{6} + \frac{1}{6} = F_X(3)$



detailed example — Discrete case:

$$F_X(a) = \sum_{x=-\infty}^{a} P_X(x)$$

$$1/6 = P_X(1) = P_X(2) = P_X(3) = P_X(4) = P_X(5) = P_X(6)$$

$$F_X(0) = 0 = F_X(-1) = F_X(-2) = \cdots$$

$$F_X(1) = P_X(1) = 1/6$$

$$F_X(2) = P_X(1) + P_X(2) = \frac{1}{6} + \frac{1}{6}$$

$$F_X(3) = P_X(1) + P_X(2) + P_X(3)$$

$\vdots$

$$F_X(6) = P_X(1) + P_X(2) + P_X(3) \ldots + P_X(6)$$

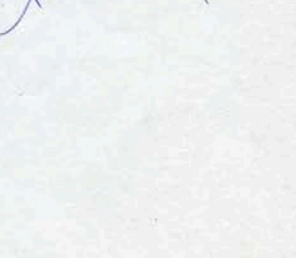
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = 1$$

$$F_X(7) = 1 = F_X(8) = F_X(9) = \cdots$$

Also

$F_X(2.5) = F_X(2)$

etc.



— Continuous case: $F_X(a) = \int_{-\infty}^{a} f_X(x)\,dx$

also: $f_X(x) = \dfrac{d F_X(x)}{dx}$

— Properties of CDF

Recall $P(X \leq a) \triangleq F_X(a)$

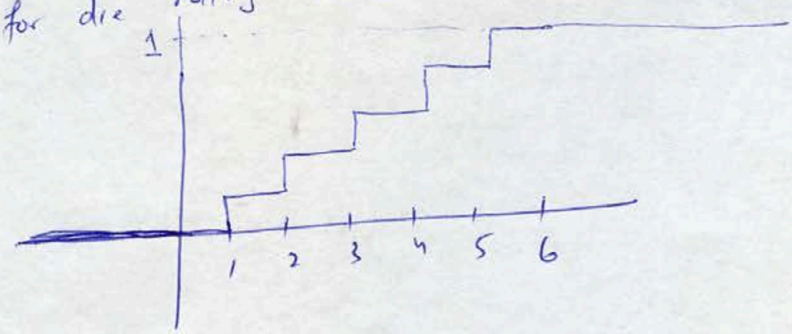
P1 $F_X(x) \geq 0 \quad -\infty < x < \infty$ (since sum of probabilities)

P2 $F_X(-\infty) = 0$

P3 $F_X(\infty) = 1$

P4 If $b > a$, then $F_X(b) - F_X(a) \geq 0$ (i.e $F_X(x)$ is a non-decreasing function) (b/c additional probability terms will be non-negative)

→ e.g for die rolling case



$F_X(x)$

Btw useful Formula: $F_X(b) - F_X(a) = P(a < X \leq b)$
(for $b > a$)

e.g die rolling
$F_X(3) = P_X(1) + P_X(2) + P_X(3)$
$F_X(5) = P_X(1) + P_X(2) + P_X(3) + P_X(4) + P_X(5)$

then $F_X(5) - F_X(3) = P_X(4) + P_X(5) = P(3 < X \leq 5)$

— continuous rv [CDF & PDF]

- For a C. rv it only makes sense to talk about probability that it lies in a range, say [a, b] b > a

$$P(a < X \leq b) = F_X(b) - F_X(a)$$

- we could make that interval rather small, say [x, x+Δ]

$$P(x < X \leq x+\Delta) = F_X(x+\Delta) - F_X(x)$$

- In fact, if the derivative of F_X exists, we may use it to define a Probability density function (PDF) for the C. rv X as

$$f_X(x) = \frac{dF_X(x)}{dx}$$

or $$f_X(x)\,dx = dF_X(x) = P(x < X \leq x + dx)$$

— so in case of cont. rv. we talk of PDF which is esentially used to find the probability that X lies in a certain range, e.g. for rang (−∞, x]

$$P[-\infty < X \leq x] = P[X \leq x] = F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$$

→ and for range (a, b] b > a

$$P(a < X \leq b) = \int_a^b f_X(x)\,dx$$

→ Properties $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ and $f_X(x) \geq 0 \;\; \forall x$

↳ since $= P(-\infty < X < \infty)$

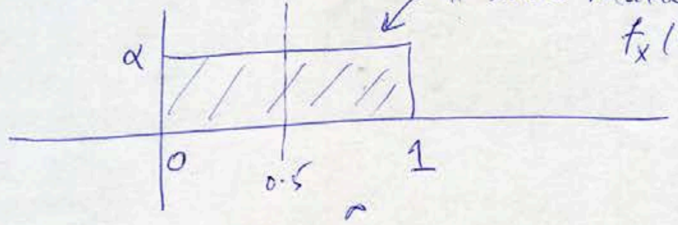
since its a derivative of a nondecreasing function ($F_X(x)$)

Tangent always positive or zero.

→ e.g $X \sim U[0,1]$ → Uniformly distributed on continuum $[0, 1]$

ie → Any value in the interval is as likely as any other (in the interval)

→ what would that relative likelihood look like?



$f_X(x) = \alpha \quad x \in [0,1]$

$\alpha = ?$ wkt $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$ ⇒ Area = 1

⇒ $(\alpha)(1) = 1$

$\alpha = 1$

→ note that $\alpha = 1$ does not represent the probability of any single point, but just the information that they are all equally likely (ie it is a 'relative likelihood')

→ what is the probability that $0.5 < X \leq 1$?

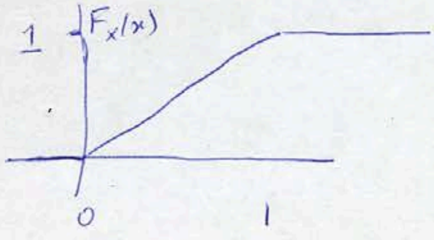
$$P(0.5 < X \leq 1) = \int_{0.5}^{1} f_X(x)\,dx = \int_{0.5}^{1} \alpha\,dx = (1)(0.5) = 0.5$$

→ i.e 50% chance that X lies in the right half of the interval → which makes sense for uniform case.

→ H.W. what is $F_X(x)$?

$\left[F_X(x) = \int_{-\infty}^{x} f_X(u)\,du \right]$

clearly
$$\begin{cases} F_X(x) = 0 & x < 0 \\ F_X(x) = 1 & x > 1 \\ F_X(x) = \int_0^x \alpha\,du & x \in [0,1] \\ \quad = \alpha x = x & \end{cases}$$



then $f_X(x) = \dfrac{dF_X(x)}{dx}$ ←

also checks out

In summary.

- so, $X \sim f_X(x)$

↑ random (continuous) — ↑ our knowledge of the relative likelihoods.

PDF

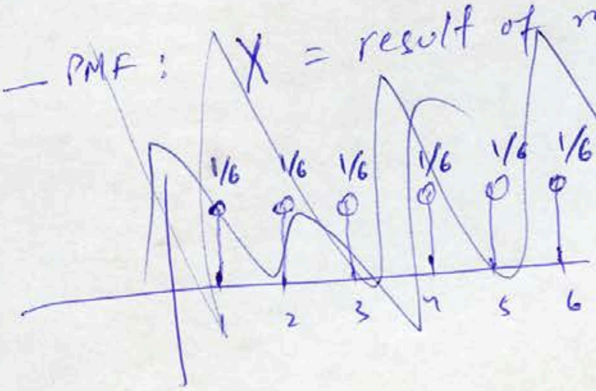
$X \sim P_X(x)$

↑ random (Discrete) — ↑ probabilities.

PMF

- the Pdf is very important as it contains all the information about the stat. of the r.v. itself. (so is PMF).

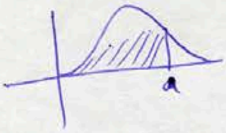
- PMF: X = result of rolling a die



Discrete case

- e.g.

Continuous



$$P(X \le a) = \int_{-\infty}^{a} f_X(x)\,dx \triangleq F_X(a)$$

$$P(a \le X \le b) = \int_{a}^{b} f_X(x)\,dx$$

(mean) $$E[X] = \int_{-\infty}^{\infty} x\, f_x(x)\,dx \triangleq m_X$$

(variance) $$V[X] = \int_{-\infty}^{\infty} (x-m_X)^2 f_X(x)\,dx$$

$$P(X=a) = 0 \quad (\text{Think why?})$$

Discrete

$$P(X \le a) = \sum_{x=-\infty}^{a} P_X(x) \triangleq F_X(a)$$

$$P(a \le X \le b) = \sum_{x=a}^{b} P_X(x)$$

$$E[X] = \sum_{x=-\infty}^{\infty} x\, P_X(x) \triangleq m_X$$

$$V[X] = \sum_{\forall x} (x-m_X)^2 P_X(x)\,dx$$

$$P(X=a) = P_X(a)$$

(3) From one to two RVs

— we may sometimes be interested in defining two random variables on a sample space Ω. E.g, rolling a die

Mapping 1

Ω

ω = number showing on the die

$X(\omega) = \omega$

$Y(\omega) = \omega^2$ $\Rightarrow$

ω	$X(\omega)$	$Y(\omega)$
1	1	1
2	2	4
3	3	9
4	4	16
5	5	25
6	6	36

Mapping 2

or

Ω

ω = number showing on the die

$X(\omega) = \omega$

$Y(\omega) = \begin{cases} 0 & \omega \text{ even} \\ 1 & \omega \text{ odd} \end{cases}$

ω	$X(\omega)$	$Y(\omega)$
1	1	1
2	2	0
3	3	1
4	4	0
5	5	1
6	6	0

Defn we say that two random variables on $(\Omega, \mathcal{F}, P)$ are ~~same~~ **equal** if they have the exact same mapping!

i.e $\boxed{X = Y}$ if $X(\omega) = Y(\omega) \;\; \forall \, \omega \in \Omega$

→ The random variables above are "surely" equal

→ Sometimes we compare rv in "almost surely" sense, such as

$X = Y$ almost surely or $X = Y$ with probability 1.

→ This kind of thing occurs generally in continuous case (range = uncountably infinite, and probability on any single value is zero)

→ e.g. let $\omega \in \mathbb{R}$ ↙ Ω, then if we say (like before)

$$X(\omega) = Y(\omega) \quad \forall \ \omega \in \mathbb{R}$$

→ then $X = Y$ surely since they are same at all values of ω

→ but what if we have

$$X(\omega) = Y(\omega) \quad \forall \ \omega \in \mathbb{R} \text{ except } \omega = 0.5 \quad \left(\begin{array}{l} \text{i.e } \forall \ \omega \in \mathbb{R} - \{0.5\} \\ \text{i.e } X(0.5) \neq Y(0.5) \end{array} \right)$$

→ i.e their mapping varies on just one (or finite?) value of ω but is the same on the rest of the continuum (infinite values)

then $P\big(\omega : X(\omega) \neq Y(\omega)\big) = 0$

alternatively $P\big(\omega : X(\omega) = Y(\omega)\big) = 1$

→ Here we say $X \overset{a.s}{=} Y$ or $X \overset{\text{Prob. }1}{=} Y$

→ similarly there are $X \overset{a.s}{\leq} Y$, $X \overset{a.s}{\geq} Y$ etc.

→ Further clarification from events

→ say you are to pick a real number randomly and uniformly from $[0, 1]$ ← Ω

→ define $A = \{0 \leq X \leq 1\}$ (i.e $A = \Omega$)

$B = \{X \text{ is negative}\}$

$C = \{0 < X \leq 1\}$ ↓ i.e $X \neq 0$

Then ① A will surely happen

② B will surely not happen

③ C will happen almost surely

note that, while $P(A) = P(C) = 1$

→ In case of C we know that there is a realization which may still occur ($X = 0$) although its assigned probability is zero (due to continuum mathematics)

→ In case of A there is no such realization.

— Bivariate Distributions (i.e. of TWO rvs)

— Just as we were interested in probability of two events happening at the same time

$$P(A \text{ and } B) \quad (\text{or } P(A \cap B))$$

— we could be interested in distributions of two random variables (defined on the sample space Ω)

Bivariate Cumulative Distribution Function

$$F_{XY}(x,y) \triangleq P(X \leq x, Y \leq y) = P[\omega : X(\omega) \leq x, Y(\omega) \leq y]$$

(eg height & weight of rent student)

— Bivariate Density function

$$f_{XY}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$$

or

$$F_{XY}(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f_{XY}(x,y)\,dx\,dy$$

} continuous case

— Bivariate Mass function

$$P_{XY}(x,y) = P(X = x, Y = y)$$

} Discrete case.

and $F_{XY}(x,y) = \sum_{x_i \leq x} \sum_{y_j \leq y} P_{XY}(x_i, y_j)$

— Independence:

- just as we saw independence of two random events as

$$P(A \cap B) = P(A) \cdot P(B)$$

- we have for independence of two random variables the following condition

$$F_{XY}(x,y) = F_X(x)\,F_Y(y) \quad \forall x,y$$

— or, we can also use Pdf of continuous r.v. to check independence as

$$f_{XY}(x,y) = f_X(x)\,f_Y(y) \quad \forall x,y$$

— or, by using PMF for discrete case

$$p_{XY}(x,y) = p_X(x)\,p_Y(y) \quad \forall x,y$$

— Some more properties

$$F_{XY}(\infty,\infty) = 1 \quad \left(\text{easy to check}\right)$$

$$F_{XY}(x,y) \geq 0 \quad \forall x, \forall y$$

$$\left.\begin{array}{l} F_{XY}(x,\infty) = F_X(x) \\ F_{XY}(\infty,y) = F_Y(y) \end{array}\right\} \text{Marginal Distributions}$$

(see more in Kobayashi ch. 3)

$$\sum_{\forall y} p_{XY}(x_i, y_j) = P_X(x)$$

$$\sum_{\forall x} p_{XY}(x, y) = P_Y(y)$$

marginal

$$\int_{-\infty}^{\infty} f_{XY}(x, v)\, dv = f_X(x)$$

$$\int_{-\infty}^{\infty} f_{XY}(u, y)\, du = f_Y(y)$$

marginal

$$\sum_{\forall x} \sum_{\forall y} p_{XY}(x, y) = 1$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(u, v)\, du\, dv = 1$$

(4) From Two → Many

→ Let $X_1, X_2, \ldots X_m$ be m rvs defined on Ω

→ Then we may define multivariate m-dimensional JD as

$$F_{\underline{X}}(x_1, x_2, \ldots, x_m) = P\left(X_1 \le x_1, X_2 \le x_2, \ldots, X_m \le x_m\right)$$

$$\underline{X} = [X_1, X_2 \ldots X_m]$$

→ $P_{\underline{X}}(x_1, x_2, \ldots, x_m)$ and $f_{\underline{X}}(x_1, x_2, \ldots, x_m)$ may be extended analogously.

→ Independence of many rvs

All the independence conditions may also be extended from two rvs to many

e.g. → RVs $X, Y, \ldots Z$ defined on Ω are independent if

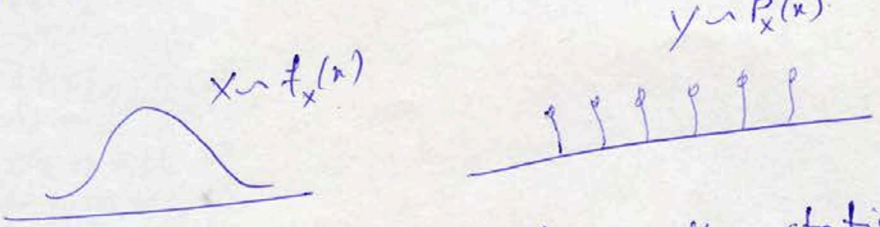
$$F_{X,Y\ldots Z}(x_k, y_l, \ldots z_m) = F_X(x_k)\, F_Y(y_l) \ldots F_Z(z_m) \quad \boxed{\forall\, x_k, y_l \ldots z_m}$$

→ PMF and PDF cases may be written analogously.

Lecture 4 (Moments)

1) From Distributions → Moments

→ we have talked in some detail about distributions
e.g PD, PMF, PDF, CDF, JPDF etc.

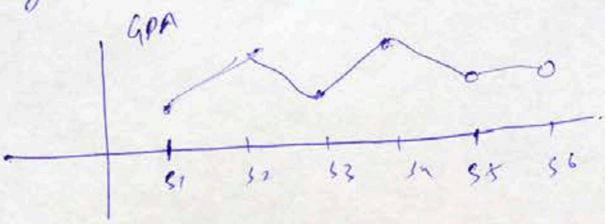


→ While distributions fully characterize the statistics of the RV, they are not always easy to work with.

→ "Too much information"

→ we often do not want all the information about things (instead some summary or overall measure is enough!)

→ e.g Ask student = "How's your performance?"



← Instead of this much detail CGPA should suffice!
↓
Single value!

→ Similarly, often instead of working with the full distributions we work with some parameters that encapsulate important properties of the distributions

→ and give useful information about the RV

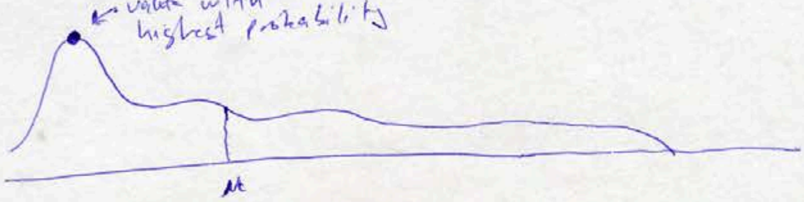
→ Among these are "Moments"

(2) So what are Moments?

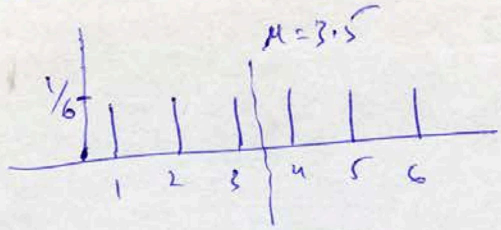
→ you are probably familiar with Mean, Variance, Covariance, central moments, non-central moments.

→ So what is "mean" (also called "Expected value")

→ Is it the most probable (most frequently occurring) value?

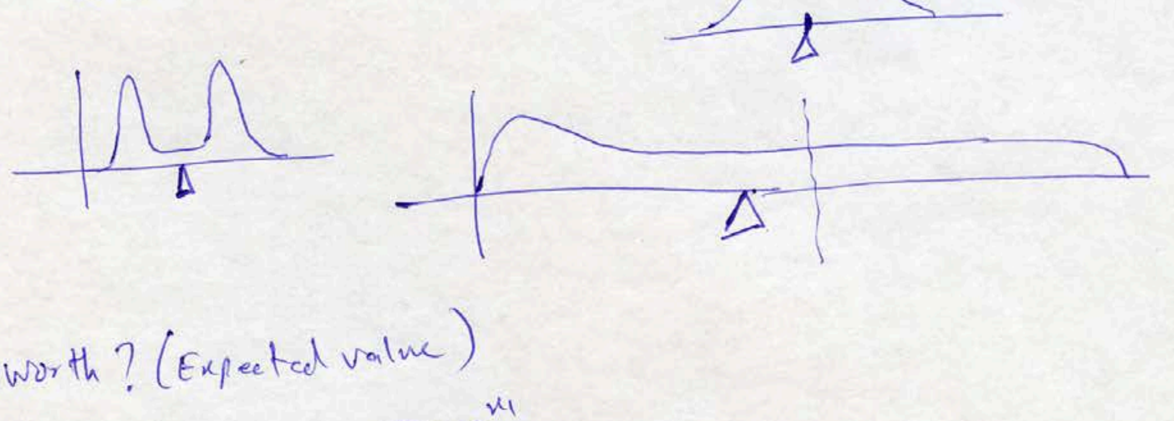


→ But let alone most probable, the mean may never even occur



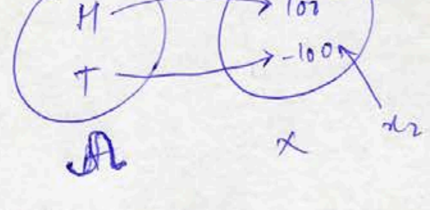
→ It is in fact the center of mass (COM) of the distribution, and the Expected worth of RV X.

COM? → Put a wedge under the distribution and see where it balances!



→ Expected worth? (Expected value)

→ let's play



fair coin. $p_x(x_1) = p_x(x_2) = 0.5$

→ How much do you expect to win or loose?

or → If you play a thousand times what net amount should you expect to win or loose per play?

$$\text{expected worth} = (0.5)(100) + (0.5)(-100)$$

$$= x_1 p(x_1) + x_2 p(x_2) \triangleq E[X]$$

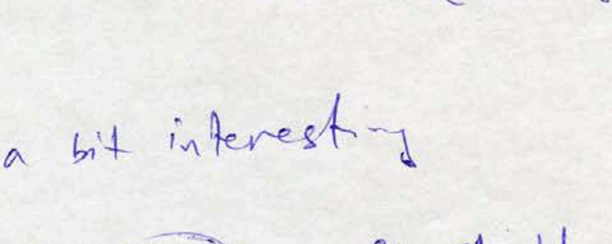
(which in this case is zero ← as expected logically)

→ e.g. in Thousand plays $\dfrac{100 \;\; 100 \;\; -100 \;\; 100 \;\; -100 \;\; -100 \;\; 100 \;\; -100 \;\ldots}{} \approx 0$

In General:

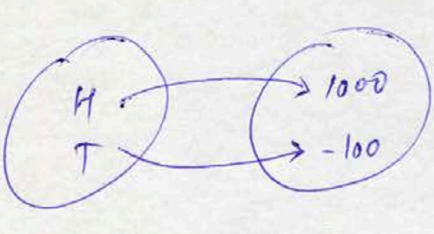
- Discrete: $E[X] \triangleq \sum_{\forall x_i} x_i P_X(x_i) = \mu_X$ ← common notation
- continuous: $E[X] \triangleq \int_{-\infty}^{\infty} x f(x)\, dx = \mu_X$

as per COM Interpretation →



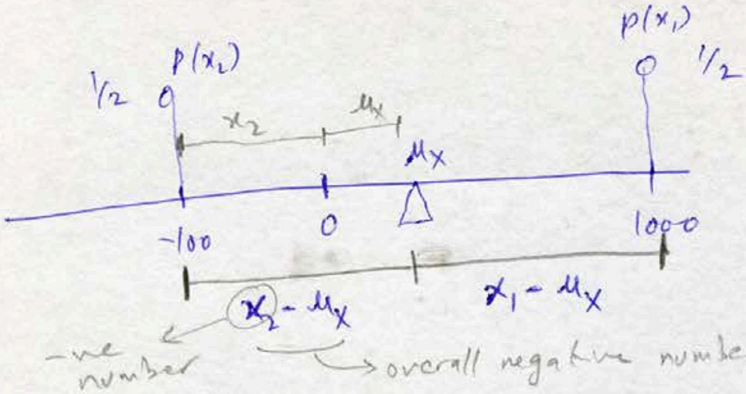
$E(X) = 0$ (Balancing point for mass)

→ Let's make it a bit interesting



Q. → Should you play?

Q. → How much expect to loose or gain on average (per play)?



where to put the wedge?

we want: $(x_1 - \mu_X) p(x_1) + (x_2 - \mu_X) p(x_2) = 0$ ← for balance

$$x_1 p(x_1) + x_2 p(x_2) - \mu_X \big(\underbrace{p(x_1) + p(x_2)}_{=1}\big) = 0$$

$$x_1 p(x_1) + x_2 p(x_2) = \mu_X$$

which is, of course, also our definition of "mean"

so here $\mu_X = (-100)(0.5) + (1000)(0.5) = -50 + 500 = 450$ ← worth per play.

→ say you play eight times, you should expect to win

$(450) \times 8 = 3600$ Rs.

(450 = worth per play; 8 = No. of plays)

→ i.e. 1000 1000 1000 1000 -100 -100 -100 -100 (more accurate as number of plays increases)

$900 \times 4 = 3600$

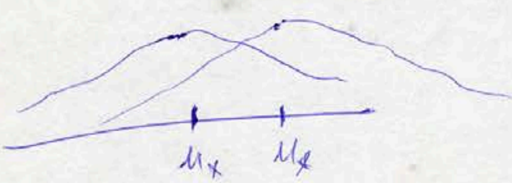
→ of course it could turn out to be all eight wins.

1000 1000 1000 1000

But that is as likely as -100 -100 -100 -100 -100 -100 . . . -100

→ So 'mean' is the center of mass, the expected worth of RV

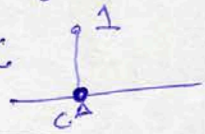
→ It kind of helps us locate the distribution (mostly!)



etc.

Q: Mean of a constant?

$E[C] = (C)(1) = C$



→ What is variance?

→ Degree of spread

→ A measure of variation (variety) in the values of the RV

→ A measure of how much spread is the RV about its mean value.

$$V(X) \triangleq E\left[(X-\mu_X)^2\right]$$

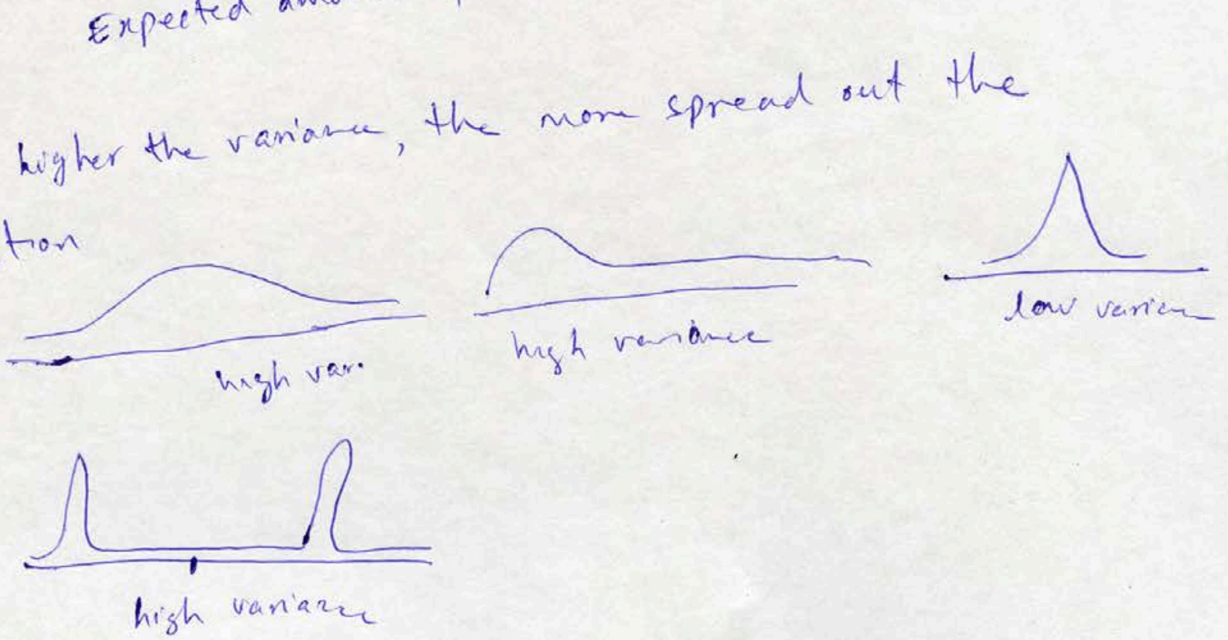
Discrete: $= \sum_{\forall x_i} (x_i - \mu_X)^2 p(x_i)$

Continuous: $= \int_{-\infty}^{\infty} (x-\mu_X)^2 f(x)\, dx$

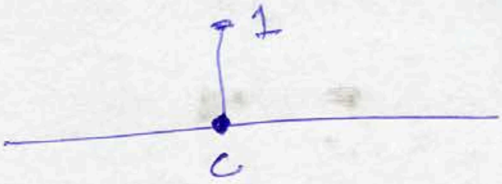
A measure of how much (in square sense) values of X vary from the mean and how often value of X vary from the mean

Expected amount of square deviation from the mean

→ In General, higher the variance, the more spread out the distribution



Q. Variance of a constant?



~~known~~ c does not deviate at all!

$$V[c] = 0$$

→ another way: $E[(C - \mu_c)^2] = \sum \overbrace{(c - \mu_c)}^{=0}{}^2 \underbrace{P(c)}_{1} = 0$, with $\mu_c = c$

(3) Some Remarks ~~& Properties~~

— we often generalize the above to define

$$E[X^k] = \sum_{\forall x_i} x_i^k P_X(x_i)$$

as the k-th (non-central) moment and

$$E[(X - \mu_x)^k] = \sum_{\forall x_i} (x_i - \mu_x)^k P_X(x_i) \quad (\mu_x \triangleq E[X])$$

(gives shape info independent of translation)

as the k-th central moment, provided they exist.

→ Continuous case defined analogously with integrals.

— clearly mean = first non-central moment (i.e. $E[X^k] = E[X]$ for $k=1$)

and variance = second central moment (i.e. $E[(X-\mu_x)^k] = E[(X-\mu_x)^2]$ for $k=2$)

— Like first and second moments (mean and variance), higher moments may also give info about shape of distribution (e.g. read about skewness ($k=3$) & kurtosis ($k=4$))

— The first and 2nd moments are often called
Low-Order statistics (mean, var, covariance) ← later!

— Third and higher moments are often called
Higher-order statistics

— These typically employ non-linear combinations of data and are harder to estimate (typically requiring more data for quality estimates)

— We often restrict ourselves to first and second moments.
— And they are often "good enough"
— In fact one of the most common distributions, Gaussian, is completely characterized by its mean and variance (more later!)
↓ — higher moments are either zero or can be expressed in terms of μ and σ alone.

— For any of the moments to exist, the relevant sum (or integral for continuous case) must converge absolutely.

— In fact, there are some distributions for which even the first moment (mean) does not exist (infinite or undefined)
— eg Pareto Distribution's special case ($\alpha \leq 1$) gives $E[x] = \infty$ ($\int x f(x) dx = \infty$)
— Cauchy Distribution has Mean undefined
$E[x]$ = Undefined
($\int x f(x) dx = \infty - \infty$) → Indeterminate (different from ∞)

— a bit more about Infinite vs. Undefined.

$$\int_{-\infty}^{\infty} x f(x)\, dx \quad \text{may be split into}$$

$$\int_{a}^{\infty} x f(x)\, dx + \int_{-\infty}^{a} x f(x)\, dx \qquad (\text{some real } a)$$

cases

$\int_a^{\infty}$	$\int_{-\infty}^{a}$	
Finite	Finite	→ Defined & finite
Finite	($\pm\infty$) Infinite	→ Defined & Infinite
($\pm\infty$) Infinite	Finite	→ Defined & Infinite
$+\infty$	$+\infty$	→ Defined & Infinite ($+\infty$)
$-\infty$	$-\infty$	→ Defined & Infinite ($-\infty$)
$+\infty$	$-\infty$	→ Undefined ($\infty - \infty$)
$-\infty$	$+\infty$	→ Undefined ($-\infty + \infty$)

— Moments: Some properties & Formulas (useful)

— Expectation is a Linear operator.

$$\therefore \quad E\left[\sum_i a_i X_i\right] = \sum_i a_i E[X_i]$$

e.g. $E[aX + bY] = a E[X] + b E[Y]$

provided RHS are well-defined

— Expected value of a function of RV X say $g(X)$

(★) — $E[g(X)] = \sum_{\forall x_i} g(x_i)\, P_X(x_i)$ → Discrete RV

or $= \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$ → continuous RV

Provided these converge absolutely

— (*) is easy to see since $Y \triangleq g(X)$ is a new RV (new mapping) with

$$y_j = g(x_i)$$

since multiple values of x may give same y.

and $P_Y(y_j) = \sum_{x_i : g(x_i) = y_j} P_X(x_i)$ ← ~~since y_j occurs only when x_i occurs for some i~~

$$\Rightarrow E[Y] = \sum_j y_j P_Y(y_j) = \sum_j y_j \sum_{x_i : g(x_i) = y_j} P_X(x_i)$$

$$= \sum_i g(x_i) P_X(x_i)$$

— Non-negativity: if $X \overset{a.s.}{\geq} 0$ then $E[X] \geq 0$

— Non-multiplicativity: In general

$$E[XY] \neq E[X]\,E[Y]$$

However if X & Y are statistically independent with finite means then

$$E[XY] = E[X]\,E[Y] \quad \text{—— (*)}$$

Proof: Recall that for independence we require

$$P_{XY}(x_i, y_j) = P_X(x_i)\,P_Y(y_j) \quad \forall x_i, y_j$$

$$\underbrace{\sum_j \sum_i x_i y_j P_{XY}(x_i, y_j)}_{E[XY]} = \sum_j \sum_i x_i y_j P_X(x_i) P_Y(y_j) \quad \forall x_i, y_j$$

$$= \left(\sum_i x_i P_X(x_i)\right)\left(\sum_j y_j P_Y(y_j)\right)$$

$$= E[X]\,E[Y]$$

— what about variance?

— $E[(X-\mu_x)^2] \triangleq V[X] \triangleq \sigma_x^2$

— $E[(X-\mu_x)^2] = E[X^2 + \mu_x^2 - 2X\mu_x] \overset{\text{by Linearity of } E[\cdot]}{=} E[X^2] + E[\mu_x^2] - 2\mu_x E[X]$

$= E[X^2] + \mu_x^2 - 2\mu_x^2$

$= E[X^2] - \mu_x^2$ ← useful formula

— Let $Z = aX + bY$, $V[Z] = ?$ given $\mu_x = E(X)$, $\mu_y = E_Y$

$\sigma_x^2 = V[X]$, $\sigma_y^2 = V[Y]$

→ note that: $\mu_Z = a\mu_x + b\mu_y$

— $V[Z] = E[(Z-\mu_Z)^2] = E[(aX + bY - a\mu_x - b\mu_y)^2]$

$= E[(a(X-\mu_x) + b(Y-\mu_y))^2]$

$= E[a^2(X-\mu_x)^2] + E[b^2(Y-\mu_y)^2] + 2E[ab(X-\mu_x)(Y-\mu_y)]$

$= a^2\sigma_x^2 + b^2\sigma_y^2 + 2ab\, \underbrace{E[(X-\mu_x)(Y-\mu_y)]}_{\triangleq\, \sigma_{x,y} \,\triangleq\, \mathrm{cov}(X,Y)}$

— so in short

$$V[aX+bY] = a^2 V[X] + b^2 V[Y] + 2ab\, \mathrm{cov}(X,Y)$$

(4) — Now what is $\mathrm{cov}(X,Y)$?

→ Degree of linear statistical relationship between X and Y

$$\operatorname{cov}(X,Y) \triangleq E\left[(X-\mu_x)(Y-\mu_y)\right] \rightarrow \boxed{\text{Normalized version: } \rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{V[X]V[Y]}}}$$

(provided finite)

↓

"on average when X deviates from its mean does Y also deviate from its mean?"

→ eg) Marks of two students (HW)

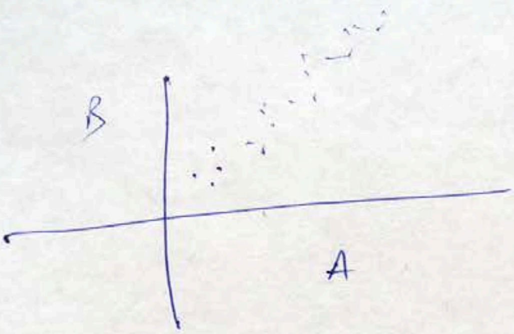
A	B
8	7
10	10
3	4
0	1
6	7
2	1
5	6
9	8

→ Seems that whenever A gets high marks, so does B (same for low)

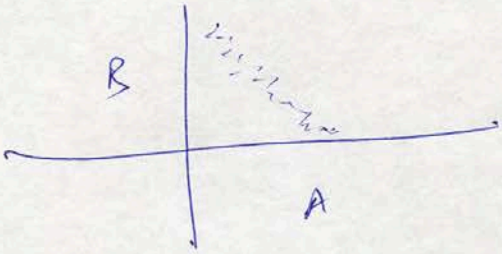
→ there seems to be some sort of linear relation here

- Coincidence?
- Copying? → (causation)
- Easy vs. Hard? (external parameter)

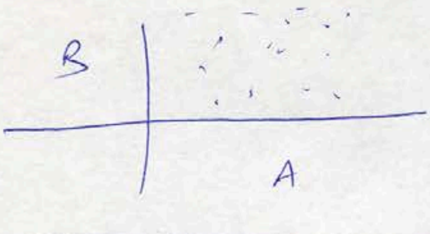
→ we say they are <u>correlated!</u>



← Both increase or decrease together
← High linear relationship
← Strong positive correlation



← As one increases other decreases
← High linear relationship
← Strong negative correlation.



← Do not vary (up or down) together
← No linear relationship
← No correlation

Remarks

cov(X,Y) ↓ ρ(X,Y) ↓

(a) — So covariance and its normalized version, correlation, give degree of linear statistical relation.

→ while covariance can have any value range correlation is always between -1 and 1 (since it is normalized)

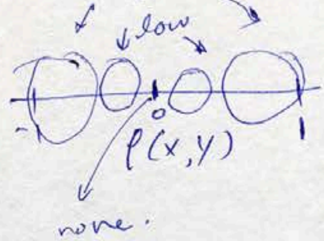
$$-1 \leq \rho(X,Y) \leq 1$$

Student got 1000 marks! ↓ out of what?

→ why normalize? Comparison easier! (just like percentages)

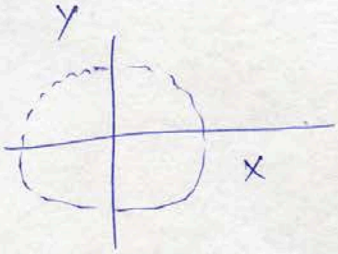
→ Large $|\rho(X,Y)|$ (i.e close to 1) means high ~~correlation~~ linear statistical relation.

→ small $|\rho(x,y)|$ (i.e close to zero) means low or none linear relationship.

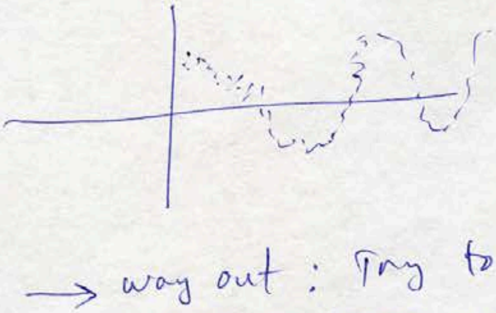


— why "linear"?

— By the way they are defined, cov or corr cannot capture non linear relationships, e.g



← $\rho(X,Y) = 0$ (even though X and Y are clearly related as they form a circle)



← $\rho(X,Y) \simeq 0$

→ way out: Try to plot the data and remove the "apparent" non-linearity first, then check corr.

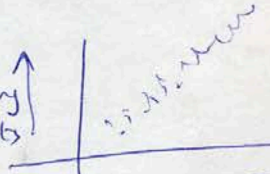
(b) Correlation Does not imply causation.

→ As we saw in student marks example, even though there was high correlation it could have come from multiple sources

high Correlation < Coincidence! / Causation? (copying?) / Hidden external variable? (easy vs. difficult?)

— A statistician has to look deeper into the problem to see which explanation is more justified.

→ e.g.



→ high correlation!

→ Does eating ice cream increases chances of drowning? (causes it?)

→ Actually, both happen mostly in summer (hidden external parameter causing both)

(c) "Correlation" is weaker than "Dependence"

Independence ⇒ Uncorrelated

Uncorrelated ⇏ Independence.

(e.g. dependence could be non-linear in which case cor[X,Y] is still zero)

i.e. Independent RVs are always uncorrelated but the converse is not always true.

— Recall that if X and Y are independent, then

$$E[XY] = E[X]E[Y] = \mu_X \mu_Y$$

$$\Rightarrow \operatorname{cov}[X,Y] = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X \mu_Y$$
$$= \mu_X \mu_Y - \mu_X \mu_Y = 0$$

So Independence $\Rightarrow$ uncorrelated (i.e $\operatorname{cov}[X,Y] = \rho[X,Y] = 0$)

~~→ Trivial example for the converse not holding.~~

~~→ Let's say $X = Y$ with same mean $\mu_X = \mu_X$ and variance $\sigma_X^2 = \sigma_Y^2$~~

~~→ then $\operatorname{cov}[X,Y] = E[XY] - \mu_X \mu_Y$~~
~~$=$~~

(d) Simpler formula → so for uncorrelated X and Y

$$V[aX+bY] = a^2 V[X] + b^2 V[Y] + 2ab \underbrace{\operatorname{cov}[X,Y]}_{=0}$$
$$= a^2 V[X] + b^2 V[Y]$$

(5) From Two → Many

→ Let $X_1, X_2, \ldots, X_n$ be random variables with finite means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots \sigma_n^2$, then the sum $S_n = X_1 + X_2 + \cdots X_n$ has

$$E[S_n] = \mu_1 + \mu_2 + \dots + \mu_n = \sum_{i=1}^{n} E[X_i]$$

and $$V[S_n] = \sum_{i=1}^{n} V[X_i] + 2\sum_{i<j} cov[X_i, X_j]$$

→ If however, all X_i are (at least) pair-wise independent

then $cov[X_i, X_j] = 0 \quad \forall i \neq j$

and we get

$$V[S_n] = \sum_{i=1}^{n} V[X_i]$$

ⓒ $C[X, K] = 0$ (K constant)

$\left(= E[(X-\mu_X)(K-\overset{K}{\mu_K})] = E[0] = 0 \right)$

→ Some useful relations.

ⓐ $Cov[X, X] = V[X]$

ⓑ $C\left[\sum_{i=1}^{K} a_i X_i, \sum_{j=1}^{l} b_j Y_j\right] = \sum_{i=1}^{K}\sum_{j=1}^{l} a_i b_j C[X_i, Y_j]$

ⓒ →

ⓓ It often helps to find variance by rewriting it as covariance (as in ⓐ above) and using ⓑ.

e.g. let $Y = X - 3$, $V(Y) = ?$

$$\Rightarrow V[Y] = C[Y, Y] = C[X-3, X-3]$$

$$= C[X, X] - C[X, 3] - C[3, X] + C[3, 3]$$

(with $C[X,3] = 0$, $C[3,X] = 0$, $C[3,3] = V[3] = 0$)

$$= V(X)$$

Lecture 5

(1) We previously saw (briefly) that

- RP = collection of random variables
- e.g. X = height of student entering ← (rv)

$$[X_1 \ X_2 \ X_3 \ X_4 \ X_5] \leftarrow \text{(rp)}$$ heights of next five students entering.

- we wrote the rp as

$$X(\omega, t) \quad t \in T$$

→ Often shortened to $X(t)$ or X_t

(with underlyiging assumption of $(\Omega, \sigma\text{-field}, P)$)

- Discrete vs. continuous
  - t discreteo, RP discrete
  - t continuous, RP continuous

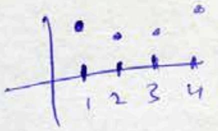
- Discrete state vs. continuous
  - values that $X(t)$ takes discrete = Discrete state RP
  - " " " " continuous = continuous state RP

- In the "Digital" world we discretize both in time and value.

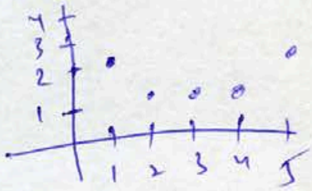
— Examples

— Temperature in the room over a day < time = continuous, state = continuous.

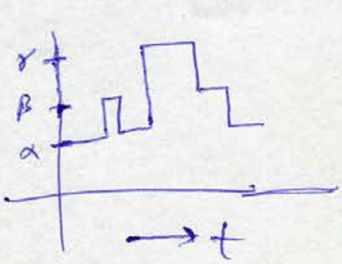
— Temperature of people in the room taken at 2 AM → continuous state, Discrete "time"



— Number of mobiles next five students entering home → { Discrete time, Discrete state } → $\triangleq$ Digital



— A voltage source that jumps randomly b/w three voltage levels < continuous time, Discrete state



(2) From Distribution ⟶ Distribution Family.

$$X \sim F_X(x) \triangleq P(X \le x) \quad \leftarrow \text{rv}$$

— RP has a <u>family</u> of distributions associated with it.

$$\{X(t), t \in T\} \sim \begin{cases} F_{X(t)}(x) \quad t \in T & \leftarrow \text{Distribution of each single } X(t) \\ F_{X(t_1), X(t_2)}(x_1, x_2) \quad t_1, t_2 \in T & \leftarrow \text{Joint Distributions of all pairs of } X(t) \\ \vdots & \\ F_{X(t_1) \cdots X(t_n)}(x_1, x_2 \cdots x_n) \quad t_1 \cdots t_n \in T & \leftarrow \text{Joint Distributions of n-tuples of } X(t) \\ \vdots & \\ \text{[upto finite dimensions]} & \end{cases}$$

— here, e.g., for $n = 4$ we have

$$F_{X(t_1),X(t_2),X(t_3)X(t_4)}(x_1, x_2, x_3, x_4) = P\left(X(t_1) \le x_1, X(t_2) \le x_2, \ldots, X(t_4) \le x_4\right)$$

→ For <u>all</u> possible values of $t_1, t_2, t_3, t_4 \in T$

— Needless to say, characterizing an RP by the entire set of distribution family can be cumbersome (exceptions exist though).

(3) Independence : Two RVs ⟶ RP

— Recall that two RVs are independent if their joint distribution is separable, i.e, X and Y are independent if

$$F_{XY}(x, y) = F_X(x)\, F_Y(y) \quad \forall x, y.$$

— To prove independence of an RP, one would have to show that every member of the distribution family is separable!

— i.e one would have to show that

$$F_{X(t_1)\cdots X(t_n)}(x_1, x_2, \ldots, x_n) = F_{X(t_1)}(x_1)\, F_{X(t_2)}(x_2) \cdots F_{X(t_n)}(x_n)$$

$\forall x_1, x_2, \ldots x_n$

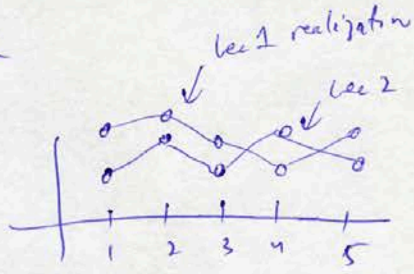
$\forall t_1, t_2, \ldots t_n \in T$

$\forall n$ $n<\infty$

— That is <u>a lot</u>! (unless simple analytical soln. exists)

(4) Moments: RV → RP

→ Height of next five students example



$$[X_1 \; X_2 \; X_3 \; X_4 \; X_5]$$

→ each one is an rv.

→ Together they form rp.

→ statistical questions we may ask!

— how is each distributed?

$$X(t) \sim F_{X(t)}(x) \qquad t = 1, \ldots, 5$$ (Distributions may be same or different)

— how are (n) of them jointly distributed?

e.g. for $n = 3$

$$[X(t_1) \; X(t_2) \; X(t_3)] \sim F_{X(t_1)X(t_2)X(t_3)}(x_1, x_2, x_3)$$

— Are they independent? (in pairs? triples? n-tuples)

— we saw that previously.

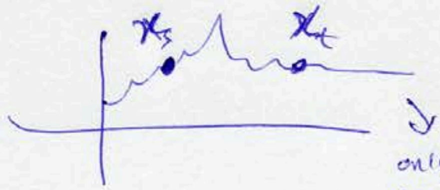
— what are the means & variances of the underlying distributions?

$$v(t) \triangleq V[X(t)] \qquad m(t) \triangleq E[X(t)]$$

(variance function) (mvf)

→ Note that we have gone from 'mean value' (for an RV) to 'mean value function' (for RP)

— Do they have some sort of statistical linear dependencies (that we can exploit)?



— $r(s,t) = C[X_s, X_t]$

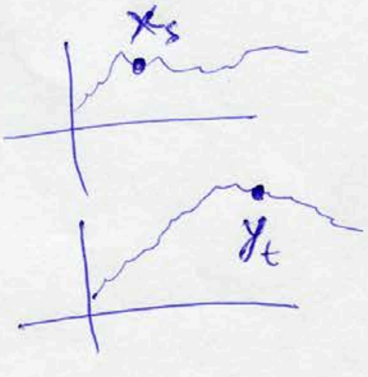
(cvf)

$$\rho(s,t) = \frac{\overbrace{C[X_s, X_t]}^{r(s,t)}}{\sqrt{V[X_s]\,V[X_t]}} \quad \begin{pmatrix}\text{normalised} \\ \text{cvf}\end{pmatrix}$$

— Do they have some linear dependencies on RVs of some other RP?

e.g $[Y_1\ Y_2\ Y_3\ Y_4\ Y_5]$ ← ages of next five students

$$r_{X,Y}(s,t) = C[X_s, Y_t] \quad \text{(ccf)}$$



— Remember "independence", "linear dependence" etc. are all informations which help us reduce our "lack of knowledge"

— Hence the interest!

Examples ⟹ RP: $[X_1\ X_2\ X_3\ X_4\ X_5\ X_6\ \dots]$

→ suppose we find (analytically or through collected data) that

$$\rho[X_1, X_2] = \rho[X_2, X_3] = \rho[X_3, X_4] \dots = \rho[X_t, X_{t+1}] = 0.8$$

and

$$\rho[X_1, X_6] = \rho[X_2, X_7] = \dots = \rho[X_t, X_{t+5}] = 0$$

→ Then we may conclude that for this RP adjacent RVs are highly correlated, whereas those apart by five time units ($\tau = 5$) are completely uncorrelated

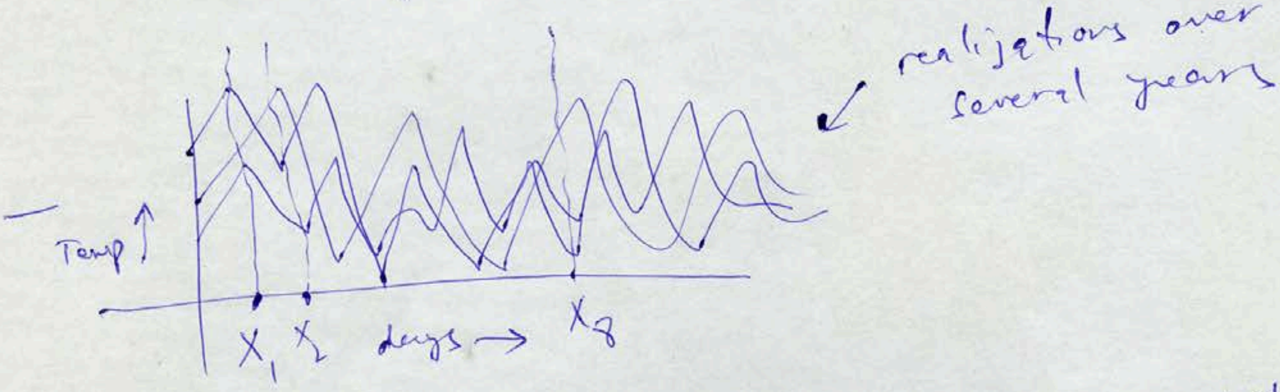
→ of course, in other processes situation may be entirely different [e.g farther points correlated but nearer ones not, etc.)

→ In general, we try to find any existing correlations and then exploit them.

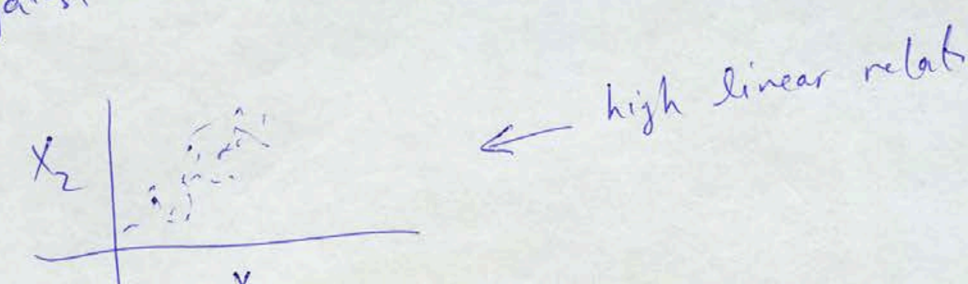
weather example:

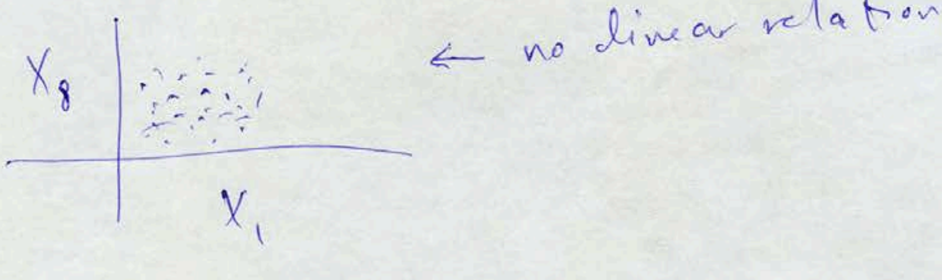
- suppose we measure temperature of a city for everyday for several years (say 10)

- Then we may be interested in finding,
  - are temps of two adjacent days correlated?
  - are days a week apart correlated?



→ we may take the ten values of X_1 and plot them against ten values of X_2 ("scatter plot")



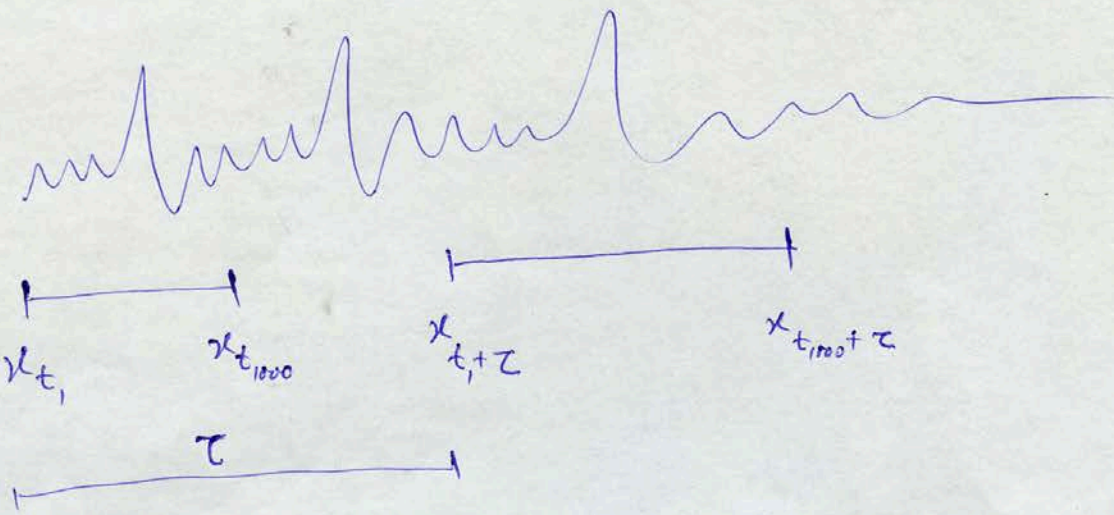


(and we may check for more pairs (X_t, X_{t+1}), (X_t, X_{t+7}))

(5) Time changes everything (well, almost everything)

— Suppose I ask two students to take 1000 samples of an industrial process on Monday morning and do statistical analysis of the data (and thereby, of the process).

— Early bird gets there early, other two hours late.



— Will they draw same statistical conclusions?

— Depends on the process!

— Do its stats change with time?

— If they don't then we don't have to worry about exactly when the data was taken (as long as enough data is taken)

— Makes life (and analysis easier)

— Such a process is statistically <u>time-invariant</u>

– Strictly stationary: RP $\{X_t, t \in T\}$ is SS if all of its statistical distributions (individual & joint) i.e, all members of its distribution family, remain unchanged after a shift in time.

– or, mathematically, RP $\{X_t, t \in T\}$ is SS if

$$F_{X_{t_1} \cdots X_{t_n}}(x_1, x_2 \cdots x_n) = F_{X_{t_1+\tau} \cdots X_{t_n+\tau}}(x_1, x_2, x_3 \cdots x_n) \quad \begin{cases} \forall n \quad n < \infty \\ \forall t_1, \ldots, t_n \in T \\ \forall \tau \quad t_i + \tau \in T \end{cases}$$

– i.e every collection $[X_{t_1}\ X_{t_2} \cdots X_{t_n}]$ for every finite (n) and every selection of $t_1, t_2 \cdots t_n \in T$ has the same multivariate distribution as the time shifted version $[X_{t_1+\tau}\ X_{t_2+\tau} \cdots X_{t_n+\tau}]$ for <u>every</u> selection of time-shift τ (such that $\tau + t_i \in T$)

– Needless to say, strict stationarity is, in general, very hard to establish.

— Moments to the rescue!

SOS ⟹ As we saw previously, full distributions can be hard to work with.

— Instead, we often work with "summarized" characteristics of distributions

— such as "moments"

→ In case of stationarity also, "strict" (or "complete") stationarity of an RP is rather complicated.

— Perhaps we can work with first & 2nd moments only and see if these are stationary?

— This leads to the concept of "second-order stationary" (SOS), weakly-stationary, or wide-sense stationary (WSS). ← Different names, same thing.

— This is often ok in practice since

— we already mostly work with first two moments

— In case of Gaussian (most common distribution)

$SOS \Rightarrow SS$ (which is not true in general)

— In general

$SS \Rightarrow SOS$

but not vice-versa (except in special cases, e.g. ↑)

— So how do we check SOS/WSS?

→ Recall the industrial process you were measuring, let's say

$$\left[X_1\ X_2\ X_3\ X_4\ X_5\ X_6 \cdots X_{100}\ X_{101}\ X_{102}\ X_{103}\ X_{104}\ X_{105}\ X_{106} \cdots\right]$$

if $E[X_1] = E[X_2] = \cdots = E[X_t] = m_x$ (i.e mean remains constant)

and $r[X_1, X_2] = r[X_2, X_3] = r[X_3, X_4] = \cdots = r[X_t, X_{t+1}]$ $\forall t$

and $r[X_1, X_3] = r[X_2, X_4] = \cdots = r[X_t, X_{t+2}]$ $\forall t$

or, in general $r[X_{t_1}, X_{t_1+\tau}] = r[X_{t_2}, X_{t_2+\tau}] = \cdots$ $\forall t$ and τ

→ i.e if the correlation between two points of the process depends only on how far apart they are (and not on their actual locations)

→ Then the two students observing

$$[X_1\ X_2\ X_3 \cdots X_{99}]$$

and $[X_{100}\ X_{101} \cdots X_{199}]$

would get the same second-order statistics!

(since then, e.g $E[X_1] = E[X_{100}]$ etc. and $r[X_1, X_2] = r[X_{100}, X_{101}]$ etc.).

— with the above in mind, it makes sense to define SOS/WSS as follows:

Defn. A random process $\{X(t), t \in T\}$ is SOS/WSS if

① It has a constant mean i.e $m(t) = m$

② Its covariance function $r(s,t)$ depends only on the time difference $\tau \triangleq t-s$ (and not on the actual values of t & s)

③ And the covariance function is finite (otherwise you may not be able to check ②)

notation ⟹ Since a WSS process's cvf depends not on actual values of s and t, but rather on their difference $t-s \triangleq \tau$, we may simplify the notation used for cvf of a WSS as follows:

→ If $X(t)$ is known to be WSS, then $\forall s, t \in T$

$$r(s,t) = c[X_s, X_t] = c[X_s, X_{s+\tau}]$$

$$= c[X_0, X_\tau] = r(0,\tau) \triangleq r(\tau)$$

→ slight abuse of notation but with the understanding that

→ Two parameters ⟹ not yet known if process WSS

→ single parameter ⟹ confirmed (or assumed) process is WSS

— Thus, we may rewrite cvf and corr. func. for WSS processes as

recall that for WSS $E[X_t] = m \quad \forall t$

$$— \; r(\tau) = C\{X_t, X_{t+\tau}\}$$
$$= E[(X_t - m)(X_{t+\tau} - m)] = E[X_t X_{t+\tau}] - m^2$$

~~and $\rho(\tau) =$~~

— This gives

$$r(0) = E[(X_t - m)^2] = V[X_t] \quad \forall t$$

— so the variance of a WSS process must also be independent of time (like the mean)

and $$\rho(\tau) = \frac{r(\tau)}{\sqrt{V[X_0]V[X_\tau]}} = \frac{r(\tau)}{\sqrt{r(0)r(0)}} = \frac{r(\tau)}{r(0)}$$

(7) Important properties of WSS cvf

→ why study
  → valuable in analysis (& analytical solns.)
  → can help check for validity of your formulation (e.g. if you claim your process is WSS, then its cvf must satisfy these properties).

→ If $r(\tau)$ is cvf of WSS $\{x(t), t \in T\}$, then

(a) $r(0) \geq 0$ (non-negative)

(b) $V[X(t+\tau) \pm X(t)] = 2(r(0) \pm r(\tau))$

(c) $r(-\tau) = r(\tau)$ (symmetric)

(d) $|r(\tau)| \leq r(0)$ (value at $\tau = 0$ sets the upper limit)

(e) if $|r(\tau)| = r(0)$ for some $\tau \neq 0$, then r is periodic.

(f) if $r(\tau)$ is continuous for $\tau = 0$, then $r(\tau)$ is continuous everywhere.

– Let's see some of the above.

(a) note that $r(0) = V[X(t)] = E[(X(t)-m)^2] \geq 0$ ← by non-negativity of expectation (if $X \geq 0$ a.s. then $E[X] \geq 0$)

(b) Recall that $V[X+Y] = V[X] + V[Y] + 2C[X,Y]$ apply to LHS and get RHS.

(c) $r(-\tau) = C[X_t, X_{t-\tau}] = C[X_{t-\tau}, X_t]$ (by symmetry of covariance)

$= r(t-(t-\tau)) = r(\tau)$

(d) From (b) wkt $r(0) \pm r(h) \geq 0$ (since variance, i.e., LHS of (b) must be non-negative)

$\Rightarrow r(0) + r(h) \geq 0 \Rightarrow r(0) \geq -r(h)$

and $r(0) - r(h) \geq 0 \Rightarrow r(0) \geq r(h)$

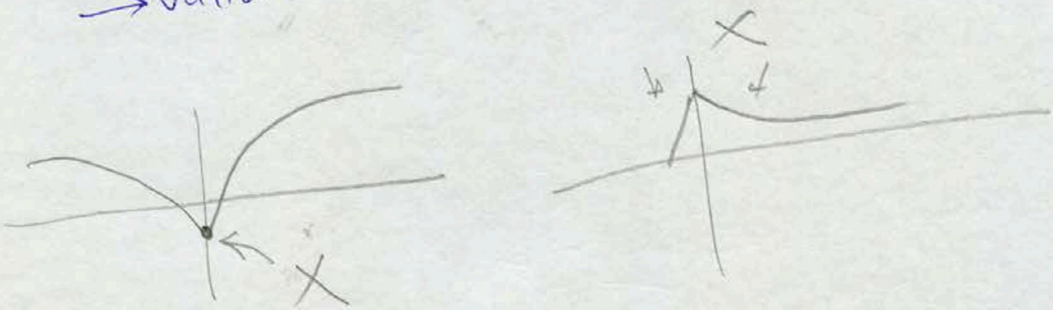
$\Rightarrow -r(0) \leq r(h) \leq r(0)$

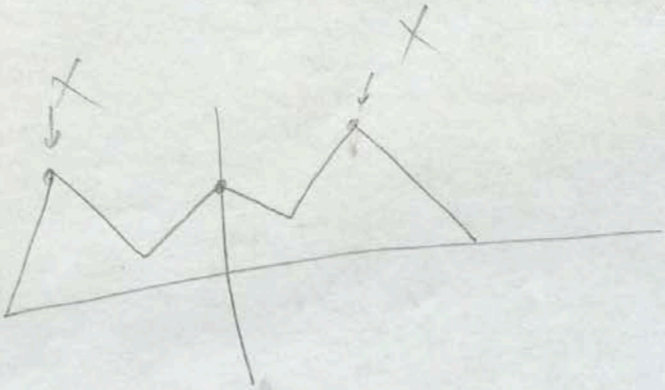
$\Rightarrow |r(h)| \leq r(0)$

(e)-(f) see [Textbook, Lindgren].

→ so what do these say:

→ valid vs invalid $r(\tau)$ for WSS





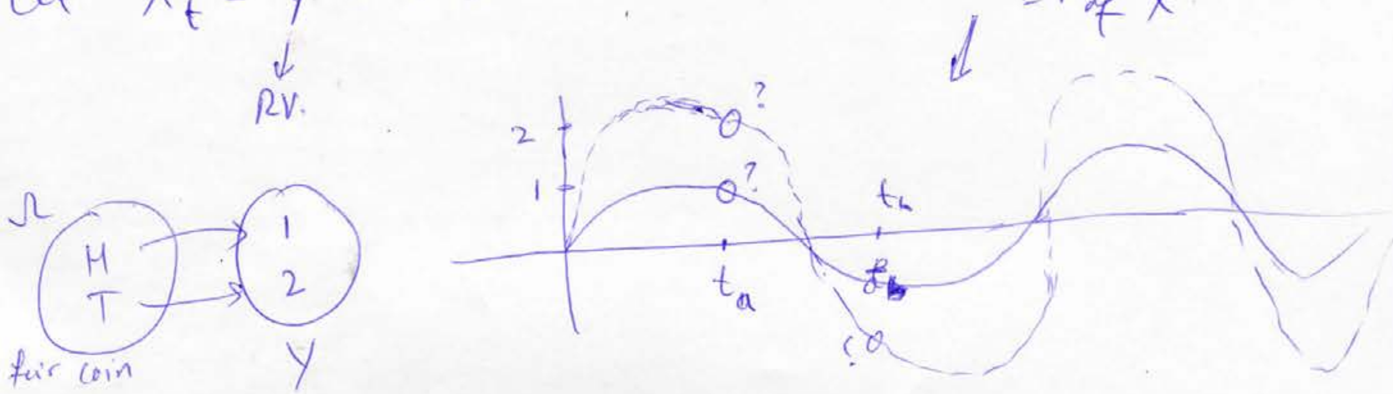
etc.

Lecture 6

(1) Let's look at some examples of RPs
— and apply some of the tools we learned.

(i) → A sinusoid with random amplitude

Let $X_t = Y \sin(t) \quad t \in \mathbb{R}$



→ Before the coin is tossed, can you tell what value X_t will take at $t = t_a$ (and $t = t_b$ etc.)?

— Interestingly, although each realization doesn't look very random, you cannot be sure of the values until one is chosen.

— Is it WSS? Assume general $Y \sim \mu_Y, \sigma_Y^2$ then

$$E[X_t] = E[Y] \underbrace{\sin(t)}_{\text{non-random}} = \mu_Y \sin(t)$$

→ the mean of X_t varies as a sinusoid (not constant)

→ Not WSS!

(ii) A "Moving Average" process.

→ consider a sequence of indep. r.v.s $\{U_t\}$ each with mean zero and variance σ^2, we may construct a new process $\{X_t\}$ by $X_t = U_t + 0.5U_{t-1}$

i.e ...... $U_{10}, U_{11}, U_{12}, U_{13}, U_{14}, U_{15}, U_{16}$ .....

$X_{10} = U_{10} + 0.5\,U_{11}$ $\quad X_{11} = U_{11} + 0.5\,U_{12}$ $\quad X_{12} = U_{12} + 0.5\,U_{13}$ etc.

→ each value of X_t is a moving (sliding) "average" of (two) U_t's.

— WSS?

also — Although U_t's are independent, are X_t's also independent of each other?

eg — Do we expect X_{10} and X_{11} to be correlated?

— how about X_{10} and X_{12}?

— etc.

— Let us check the mean

$$m(t) \triangleq E[X(t)] = E\left[U_t + 0.5\,U_{t-1}\right] = E[U_t] + 0.5\,E[U_{t-1}] = 0$$

→ mean is constant ✓

— Does covariance $r(s,t)$ depend only on $|s-t|$?

— $r(s,t) = C[X_s, X_t] = C[U_s + 0.5U_{s-1}, U_t + 0.5U_{t-1}]$

$= C[U_s, U_t] + 0.5C[U_s, U_{t-1}] + 0.5C[U_{s-1}, U_t] + 0.5^2C[U_{s-1}, U_{t-1}]$

→ hard to solve in general, so let's check for a few values

→ let $s = t+1$ (i.e $s-t=1$)

$\Rightarrow r(t+1,t) = C[U_{t+1}, U_t] + 0.5C[U_{t+1}, U_{t-1}] + 0.5C[U_t, U_t] + 0.5^2C[U_t, U_{t-1}]$

$= 0 + 0 + 0.5\sigma^2 + 0 = 0.5\sigma^2$

→ Try for $s = t-1$ (i.e $t-s=1$), and you will find

$r(t-1,t) = 0.5\sigma^2$

⇒ try with $s=t$, and you will find

$r(t,t) = 1.25\sigma^2$

⇒ try with $|s-t| \geq 2$, e.g $s = t+2$ etc.

$r(t+2,t) = C[U_{t+2}, U_t] + 0.5C[U_{t+2}, U_{t-1}] + 0.5C[U_{t+1}, U_t] + 0.5^2C[U_{t+1}, U_{t-1}]$

$= 0 + 0 + 0 + 0 = 0$

and so on for $r(t+3,t)$, $r(t+4,t)$ and $r(t-2,t)$ etc.

— so, in conclusion, we have

$$r(s,t) = \begin{cases} 1.25\sigma^2 & |s-t| = 0 \\ 0.5\sigma^2 & |s-t| = 1 \\ 0 & |s-t| \geq 2 \end{cases}$$

→ i.e values of $r(s,t)$ depends only on difference $|s-t|$ and not on individual values of s & t

— ∴ X_t is WSS (← mean constant, covariance depends only on time lag, variance finite)

— so, as before, we can now simplify the notation as

⇒ let $\tau = s-t$ then,

$$r(\tau) = \begin{cases} 1.25\sigma^2 & \tau = 0 \\ 0.5\sigma^2 & |\tau| = 1 \\ 0 & |\tau| \geq 2 \end{cases}$$

→ note that this makes sense from the "sliding" window view of X_t as well.

→ In fact, as we will see later, all Moving average processes have this property that their covariance drops to zero after a while. (i.e only nearby values are correlated).

— Interesting digression:
  — How is "mean" (expected value) related to the every day notion of "average"?

e.g. what is the average of 4 and 5?

— 4.5, how did you get that?

$$\frac{4+5}{2}$$

— How is this related to our definition of mean (as expected worth or com)?

$$E[x] = \sum_{\forall x_i} x_i \cdot P_X(x_i)$$

— Actually when given a sequence of numbers and asked to find their average (with no additional information!!) we simply assume that each of them are equally likely!

→ so $P(4) = \frac{1}{2}$ and $P(5) = \frac{1}{2}$

$$\mu = (4)(\tfrac{1}{2}) + (5)(\tfrac{1}{2}) = \frac{4+5}{2} = 4.5$$

→ What if I had given you 4, 5, 5?

$$\Rightarrow \frac{4+5+5}{3} = \frac{14}{3} \approx 4.66$$

→ then you simply assumed that (5) was twice as likely as (4)

$$(4)\left(\frac{1}{3}\right)+(5)\left(\frac{2}{3}\right)=\frac{4+10}{3}=\frac{14}{3}$$

→ so basically, "average" is our "best effort" in the light of what information is available

(iii) A random harmonic (sinusoidal) oscillation

Let $X(t) = A\cos(2\pi f_0 t + \phi)$ $f_0 > 0$ (fixed)

→ where A and ϕ are RVs independent of each other, and

$$\phi \sim U[0, 2\pi]$$

$A > 0$ with finite variance.

→ Is the process WSS?

→ recall that we found $A\sin(t)$ to be not WSS

→ but here the phase is also shifted randomly

→ Let's check.

$$E[X(t)] = E[A\cos(2\pi f_0 t + \phi)] = E[A]\cdot E[\cos(2\pi f_0 t + \phi)]$$ (by indep of A & ϕ)

here → $E[\cos(2\pi f_0 t + \phi)] = \int_{-\infty}^{\infty} \cos(2\pi f_0 t + y)\, f_\phi(y)\, dy$

$$= \int_0^{2\pi} \cos(2\pi f_0 t + y)\frac{1}{2\pi}\, dy = 0$$ (e.g. expand integral and see)

→ And the covariance is found ~~to be~~ as
(recall that $X(t)$ is zero mean.)

$$E\left[X(s)X(t)\right] = E\left[A^2\cos(2\pi f_0 s+\phi)\cos(2\pi f_0 t+\phi)\right]$$

$$= E\left[A^2\right] \cdot E\left[\cos(2\pi f_0 s+\phi)\cos(2\pi f_0 t+\phi)\right]$$

(indep. of A & ϕ)

← see [Book, Lindgren]

$$= \frac{1}{2}E\left[A^2\right]\cos(2\pi f_0 (s-t))$$

(also give $V[X_t]$, i.e. with $s=t$: $V[X_t] = \frac{1}{2}E[A^2]$)

→ i.e. covariance only depends on the difference $|s-t|$

(recall that $\cos(-\theta) = \cos(\theta)$ also).

→ the process is therefore WSS (← constant mean; cov. depends only on lag; finite variance)

→ An important thing to note here is that the covariance function is also harmonic with the same frequency (f_0), except only that the random amplitude is now replaced by variance of the random amplitude.

(iii) → In fact, it can be shown that a random process made up as a superposition of sinusoids

$$X(t) = A_0 + \sum_{k=1}^{n} A_k\cos(2\pi f_k t + \phi_k)$$

with each $\phi_k \sim U[0, 2\pi]$ and all A_k's with finite variances (say $E[A_k^2] = \sigma_k^2$) and all ϕ's and A's independent of each other will have a covariance function

$$r(\tau) = \sigma_0^2 + \sum_{k=1}^{n} \frac{\sigma_k^2}{2} \cos(2\pi f_k \tau)$$

→ i.e. the covariance function is also a superposition of sinusoids of frequencies f_k

→ This fact will prove helpful when finding/defining Fourier Tx of RPs!

iv DTMC

consider a sequence of random variables $\{X_k\}$

$$\ldots\ldots X_{10}\ X_{11}\ X_{12}\ X_{13} \ldots\ldots X_{n-1},\ X_n,\ X_{n+1} \ldots\ldots$$

such that

- Each X_k is discrete in state and takes states from $S = \{0, 1, 2, \ldots\}$ ← "state labels"

- $P[X_n | X_{n-1}, X_{n-2}, \ldots] = P[X_n | X_{n-1}]$ $\forall n$.

→ That is: future value (X_n) depends on the past and current values $(X_{n-1}, X_{n-2}, \ldots)$ only through the current value (X_{n-1})

→ i.e knowing X_{n-1} alone gives the same information about X_n as knowing all the past values of the sequence.

→ This very simple assumption ~~idea~~, called "Markov Property" (also "memoryless property") forms the basis of an entire field of statistics ⟹ Markov chains!

→ Finds application in physics, economy, operations research, literature, biology, machine learning ...

→ Markov himself applied it to Pushkin's long poem "Eugene Onegin" to predict probability of a vowel appearing after a vowel etc.

→ The sequence described above is called a "Simple Markov Chain" or a "First-order Markov chain".

(v) We can similarly define a hth order Markov chain as

$$P\left[X_n \mid X_k \quad -\infty < k \leq n-1\right] = P\left[X_n \mid X_{n-1}, X_{n-2}, \dots X_{n-h}\right] \; \forall n$$

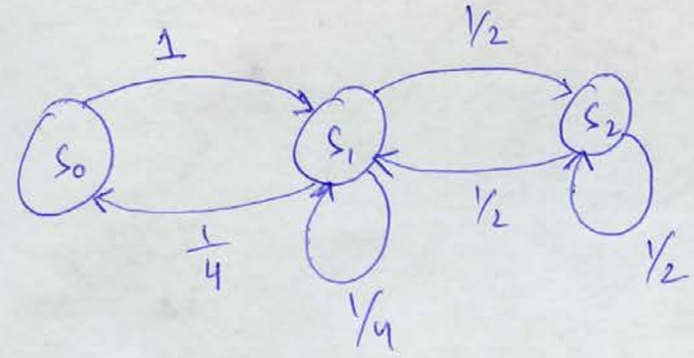
where each X_n takes discrete states labeled $S = \{0, 1, 2, \dots\}$

→ e.g for 2nd order Markov chain, we have

$$P[X_n | X_k \; -\infty < k \le n-1] = P[X_n | X_{n-1}, X_{n-2}] \quad \forall n$$

→ i.e knowing last two values is same as knowing all past values as far as predicting X_n is concerned.

Visualizing → We often visualize Markov chains through "State Transition Diagrams"



called "Transition Probabilities".

e.g

$$P[X_n = 1 | X_{n-1} = 0] = 1 \triangleq P_{01}(n)$$

(1 ← S_1, 0 ← S_0)

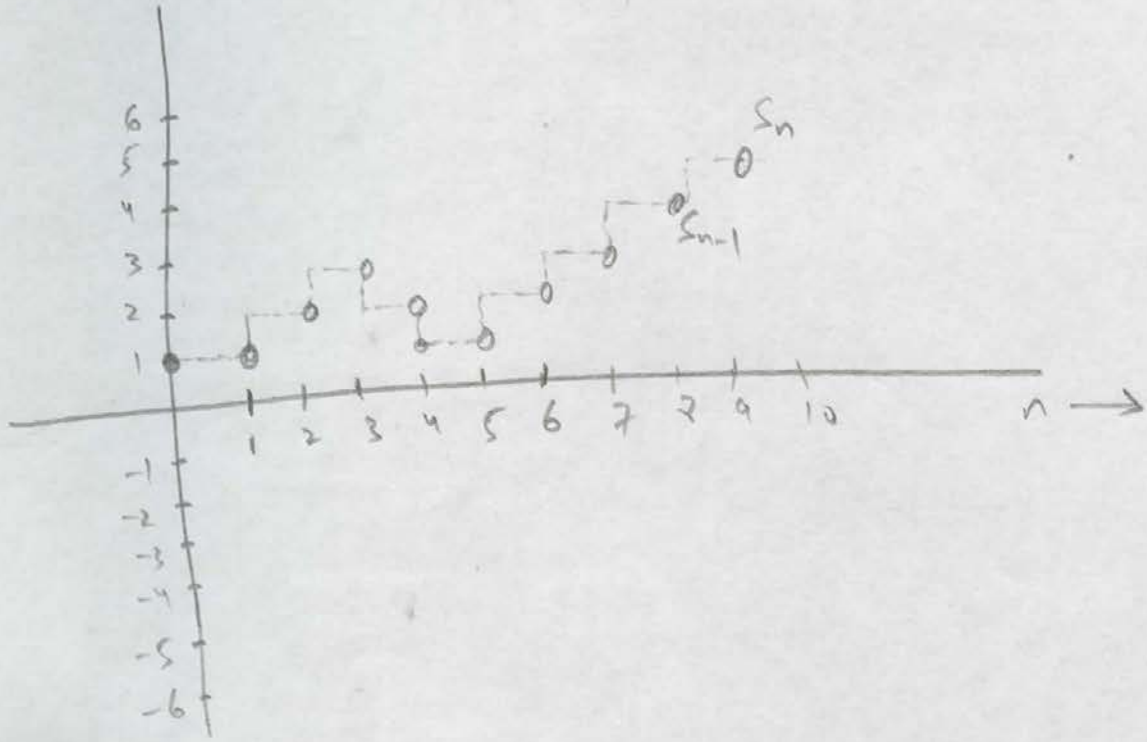
$$P[X_n = 1 | X_{n-1} = 1] = 1/4 \triangleq P_{11}(n)$$

(vi) Random Walk (a Markov chain)

→ Consider a DT-DS RP where at each step you either go up by one step (+1) with probability P, or down by one step (-1) with probability 1-P, such that each choice b/w +1 or -1

(11)

is independent of the past choices. Then a particular realization of such a process may look like this



→ Then we may write its current state S_n as

$$S_n = S_{n-1} + X_n \qquad \left(\text{or } S_n = \overbrace{X_1 + X_2 + \dots + X_{n-1}}^{S_{n-1}} + X_n\right)$$

→ where $P[X_n = +1] = p$ and $P[X_n = -1] = 1 - p \quad \forall n$

and $X_1, X_2, \dots, X_n$ are independent

Terminology → Such RVs that are < Independent of each other / And all follow the same distributions

are called "i.i.d" (Independent and Identically Distributed)

→ Clearly, state S_n depends on the past states $(S_{n-1}, S_{n-2}, S_{n-3} \dots)$ only through S_{n-1}

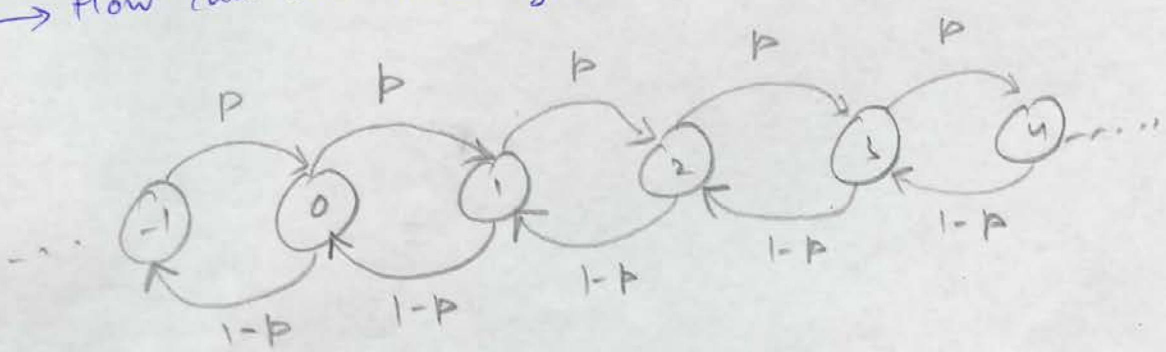
i.e $P[S_n \mid S_{n-1}, S_{n-2}, \cdots] = P[S_n \mid S_{n-1}]$

→ i.e, S_n is a Markov chain, and

$$P[S_n = i+1 \mid S_{n-1} = i] = p \quad (\forall i \in \mathbb{Z})$$

and $$P[S_n = i-1 \mid S_{n-1} = i] = 1-p \quad (\forall i \in \mathbb{Z}$$

→ How can we visualize this chain?



→ In fact, the process we just described is a very important type of Markov chain called "Random Walk".

└→ Extremely useful in statistical modeling.

→ Note that a random walk is not WSS, since

$$E[S_n] = n(p-q) \quad (q \triangleq 1-p)$$
$$(= E[X_1] + E[X_2] + \cdots E[X_n])$$

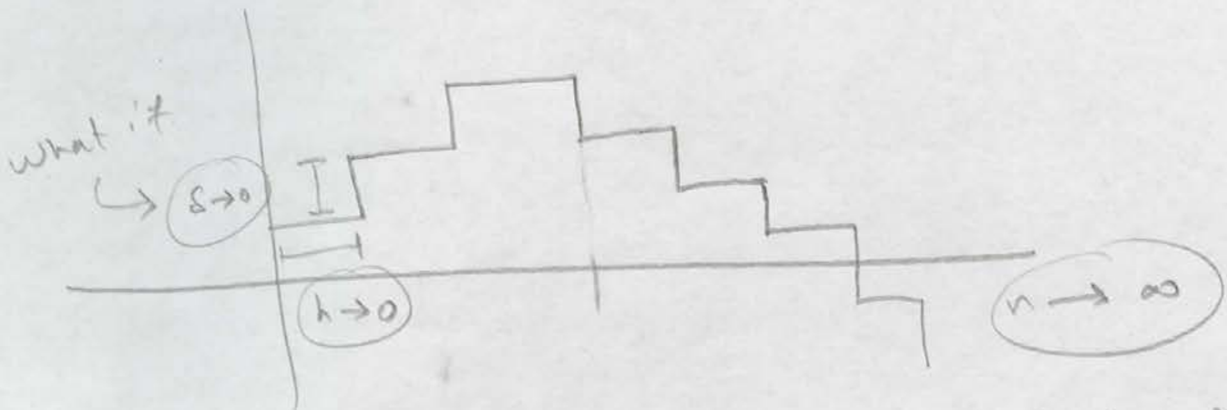
and $$R(m,n) = mn(p-q)^2 + 4pqn \quad m \geq n$$

} see, e.g. Kobayashi ch#17

(VII) Brownian Motion / Wiener Process

→ we discussed random walk as a DT-DS Markov process.

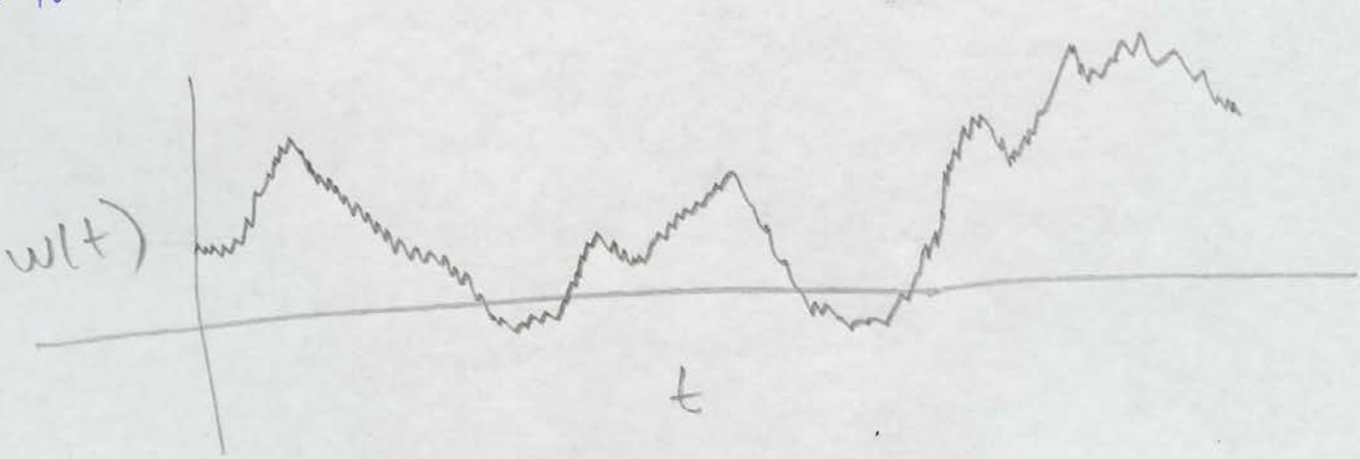
→ what if we allowed both time and state to become continuous (+ some additional conditions)



such that $\frac{\delta^2}{h} \to \alpha$ ($\alpha > 0$ some constant)

↓ Additional condition to stop process from, e.g., degenerating

→ Then the process may look like (one realization)



Mathematically we say that the limiting process $W(t)$ such that

$$\lim_{\substack{n\to\infty,\, h\to 0\\ \delta\to 0}} S_n = W(t) \quad \text{with } t = nh,\ \frac{\delta^2}{h} = \alpha$$

(and where $X_i = \begin{cases} +\delta & \text{with probability } 1/2 \\ -\delta & \text{with probability } 1/2 \end{cases}$)

(recall, $S_n = X_1 + X_2 + \dots + X_n = S_{n-1} + X_n$)

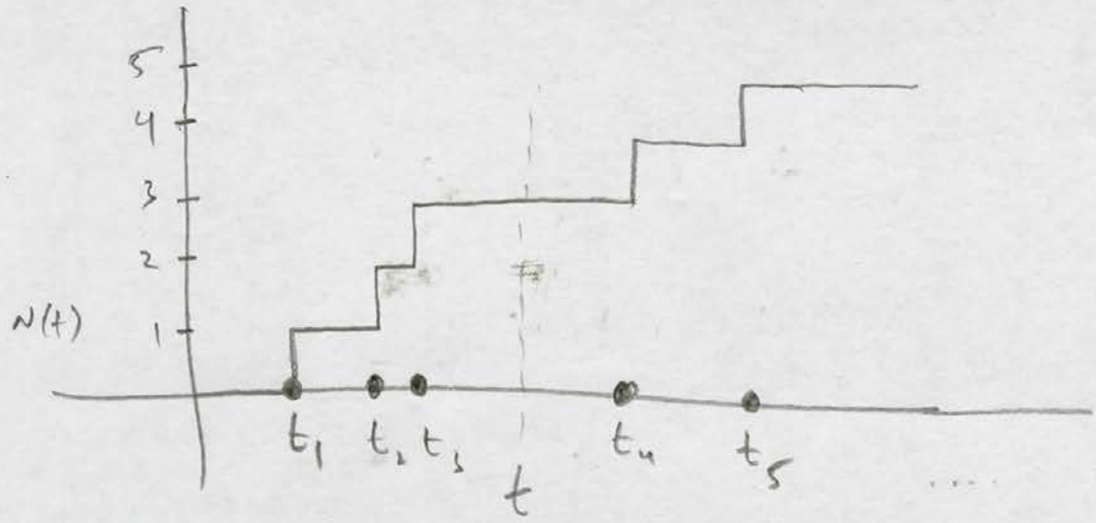
→ Is a Brownian Motion (also called "Wiener process")

→ In fact, Brownian Motion is a special case of a much larger class of CT-CS Markov processes called "Diffusion processes".

(Interesting: $W(t)$ follows Gaussian Distribution) ← more later.

(viii) Point & Counting Processes

— consider a CT-DS process where "point events" (e.g birth, call, customer arrival ...) occur randomly at $t_1, t_2, \dots t_n$ and we keep a count of the cumulative events in $N(t)$ = count of events in $(0, t]$



→ Note that we are interested in "point events"

→ so such random processes are often called "point processes"

→ $N(t)$ keeps a "count" of events that have occurred so far (in $(0, t]$) and is random.

→ so we often call such RP as a "counting process"

(IX) Renewal Processes (& special case: Poisson Process)

→ Let us talk about the random spacing b/w occurance of events in the point process, i.e, let us talk about

$$X_n \triangleq t_n - t_{n-1}$$

→ If the spacings $\{X_n\}$ are i.i.d RVs (with common distribution $F_X(x)$) then we get a special class of point (or counting) processes called "Renewal Processes" → i.e $N(t)$ = Renewal Process if $\{X_n\} \sim$ i.i.d

→ Furthermore; if the common distribution of the i.i.d $\{X_n\}$ is ~~Poisson~~ Exponential (i.e $F_X(x) \sim$ ~~Poisson~~ Exponential) then we get a special sub-class of Renewal processes called "Poisson Process".

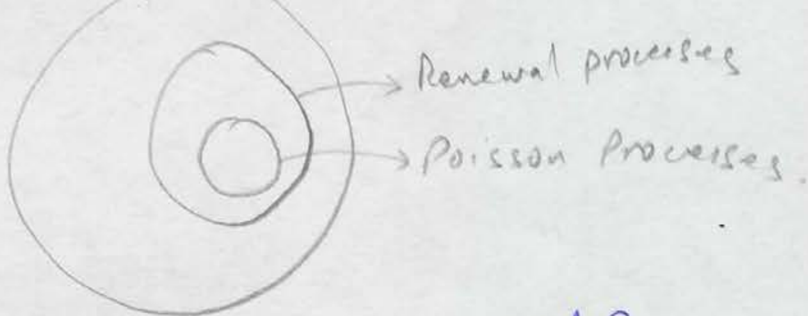
→ why "Poisson"?

→ it can be shown that if $\{X_n\}$ are i.i.d with $F_X(x) \sim$ Exponential then

$N(t) \sim$ Poisson

(more later.)

→ Thus we have:



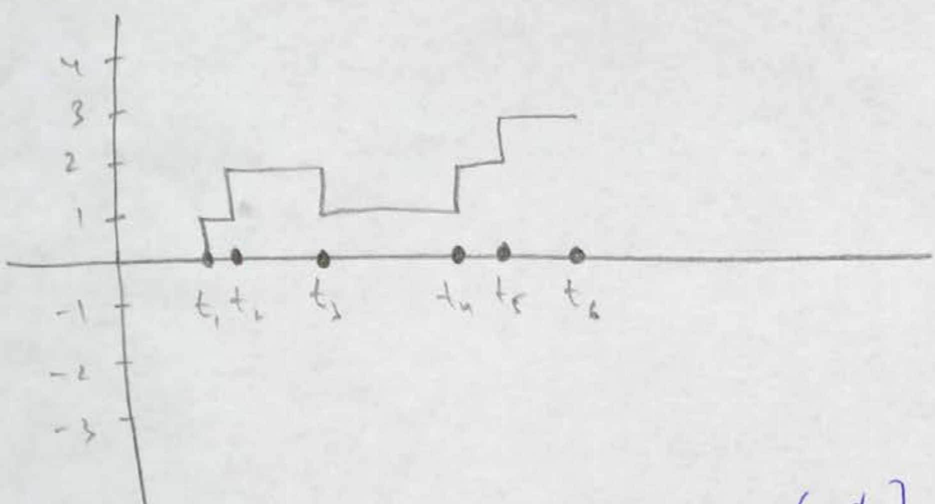
→ In fact a Poisson process is also a CTMC.
(Continuous-time Discrete-state Markov chain).

→ CTMC requires spacings to have exponential distribution (otherwise the resulting process fails to satisfy Memoryless (Markovian property and is called a Semi-Markov process).

(X) Birth-Death Processes

→ A poisson process is a "pure Birth" process, i.e., it only counts "arrivals" but no "departures". It is thus non-decreasing.

→ In many applications we are interested in counting events that represent both births (arrivals) and death (departures) — e.g. population studies, bacteria evolution, queueing theory, net accumulation models etc.



$$N(t) = B(t) - D(t)$$

- $B(t)$: count of Births in $(0, t]$
- $D(t)$: count of Deaths in $(0, t]$
- $N(t)$: Net "alive" population.

→ $N(t)$ is a Birth-Death process:

Lecture 7

(1) Expectation (Ideal) → Reality (Practical).

- we have talked about random variables and deriving something "known" (loosely "deterministic") from their statistics, e.g., moments (mean, variance ...)

- Similarly, we've done something analogous for random processes as

"fixed"

random process → Mean value function (mvf)
covariance function (cvf)
etc.

- And in so doing, we have defined some exact formulae, e.g.,

r.p. ⇒ $[X_1\ X_2\ X_3\ X_4\ X_5\ X_6 \cdots X_n \cdots]$

X_1 ↓

$x_{11} \bullet p(x_{11})$
$x_{12} \bullet p(x_{12})$
$x_{13} \bullet p(x_{13})$
$x_{14} \bullet p(x_{14})$

X_n ↓

$x_{n1} \bullet p(x_{n1})$
$x_{n2} \bullet p(x_{n2})$
$x_{n3} \bullet p(x_{n3})$
⋮

→ with mvf defined nicely as $E[X_n] = \sum_{\forall i} x_{ni}\, p(x_{ni}) \quad \forall n$

→ note that unless the process is stationary, each X_n may have a different mean (and sample space and PMF ...)

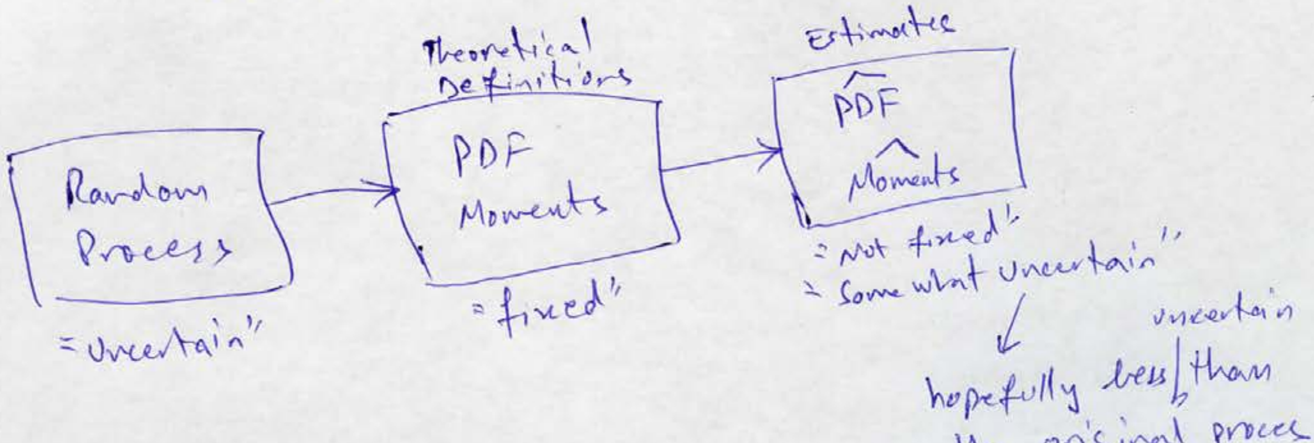
→ The above is all nice for [simple] theoretical cases. But reality is often different.

→ In reality, we may not know the underlying distributions of members of $\{X_n\}$ (or even their sample spaces).

→ Q. How do we then calculate (i.e "estimate") the statistics of the samples (r.v.) of an r.p.?

→ Interesting: note that since an "estimate" may be "good" or "bad", and estimates of the same thing may vary (between methods and data collected), we are again going towards something not entirely "fixed" (deterministic)

→ i.e., we have gone back to something somewhat "random"



→ Today's lecture mostly about these "estimates"
— How to find?
— When feasible?
— How to judge? (good or bad!)

(2) The problem: In practice often only process "observation" (data) is available

— and NOT the knowledge of its underlying random phenomena.

— How to find statistics? (e.g. moments?)

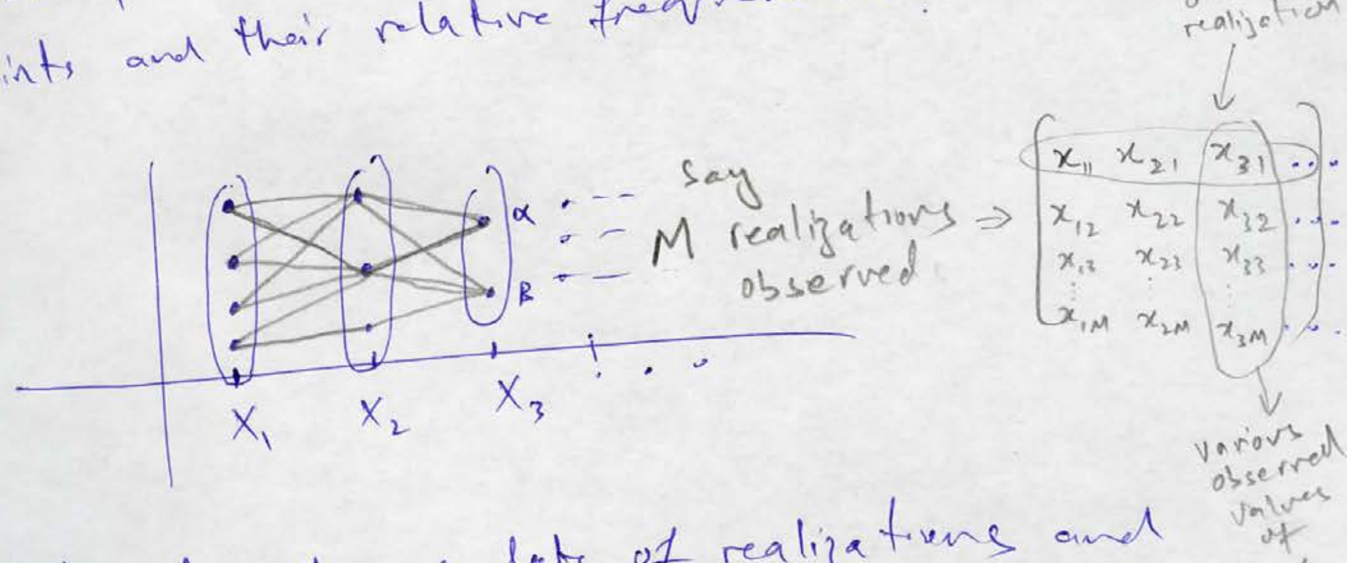
— e.g. $[X_1\ X_2\ X_3 \ldots]$ r.p.

→ each r.v. has underlying distribution (and sample space)

→ But I don't know it.

→ How can I estimate?

→ Perhaps observe lots of realizations so that each of X_1, X_2 and X_3 show all their sample space points and their relative frequencies?



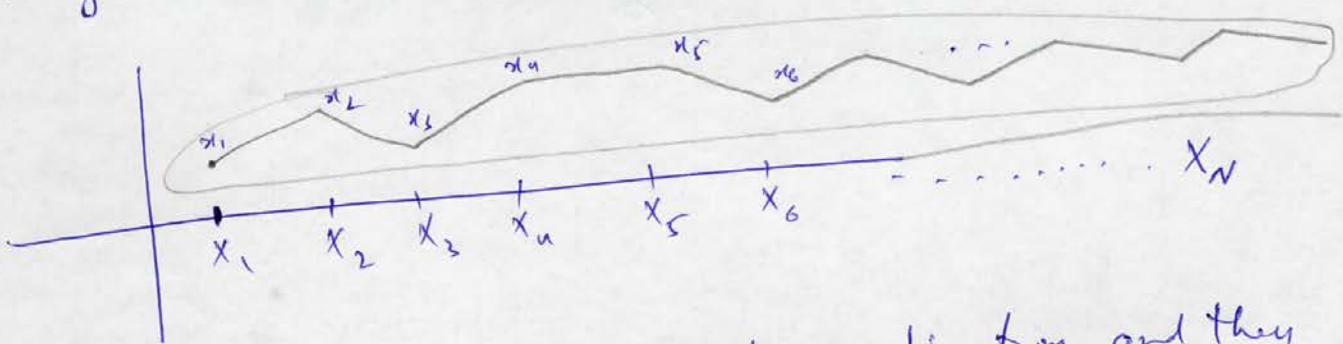
→ Strategy 1: observe lots of realizations and then estimate stats (e.g means) of X_1, X_2, X_3 from these

# of times α observed (out of M)

e.g. $\widehat{E[X_3]} = \frac{\sum_{i=1}^{M} x_{3i}}{M} = \frac{K_\alpha \alpha + K_\beta \beta}{M}$ ← "Ensemble" Average at $t=3$

Strategy 2

→ observe one very long realization of the r.p. and try to estimate stats (e.g. mean) from that long-time realization.



→ e.g. observe N samples of a single realization and then estimate

$$\widehat{E[X]} = \frac{\sum_{i=1}^{N} x_i}{N} \leftarrow \text{time average of the process.}$$

→ Q. would the second strategy work?

- It is clearly simpler than first one.
- one pre-condition is pretty obvious, for a single value (time mean) to represent the ensemble mean at all time points we must have

$$E[X_1] = E(X_2) = E[X_3] \ldots$$

- i.e. mean should be stationary (constant)
- otherwise a single $E[X]$ can never represent them!

— what else would be required?

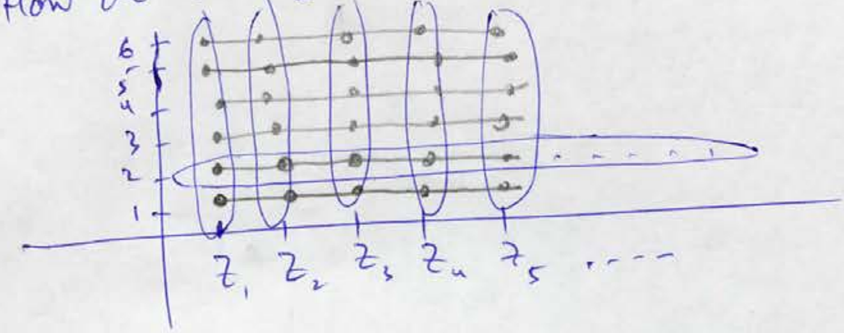
— let's see:

Example 1 Throw a single die and define a r.p. as

$\{Z_n\}$ such that each $Z_n = \omega$ where

ω = outcome of throwing the die the first time.

→ How do realizations of such a process look?



→ ensemble mean at each time is 3.5

$$E[Z_1] = E[Z_2] = \dots = E[Z_N] = 3.5$$

→ mean is stationary.

→ However, you cannot get this mean from a single realization (no matter how long!)

→ e.g. each single realization gives a different mean

→ for depicted one $E[Z] = 2$ etc.

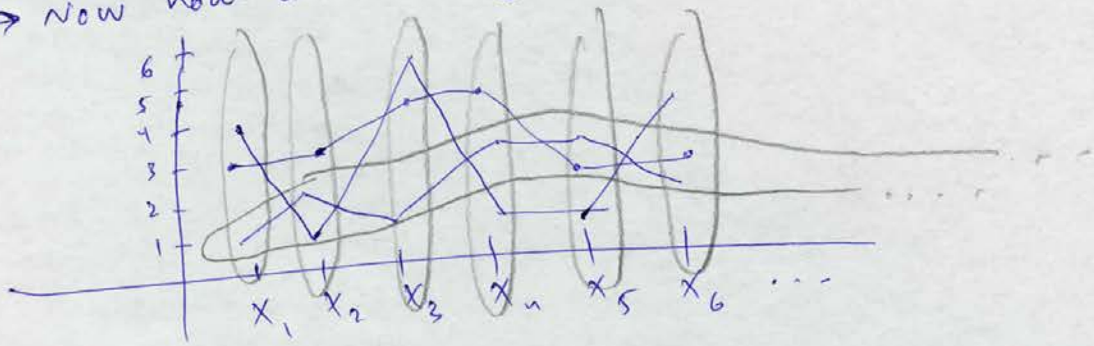
Example 2

Define a random process $\{X_n\}$ such that

$X_n = \omega_n$ where ω_n is the outcome

of nth independent throw of a die.

→ Now how do realizations look?



→ Ensemble mean at each time is

$$E[X_1] = E[X_2] \cdots = 3.5$$

→ Easy to guess that if we observe a single realization of this process for pretty long time, we should get the same mean (i.e. throw a die thousands of times and see the mean), which is

$$E[X] = 3.5$$

→ Both processes (EX1 & EX2) are stationary, still in one of them ensemble mean = time mean (ex.2)

while in the other ensemble mean ≠ time mean (ex.1)

→ Alexandar Dumas.

Three Musketeers

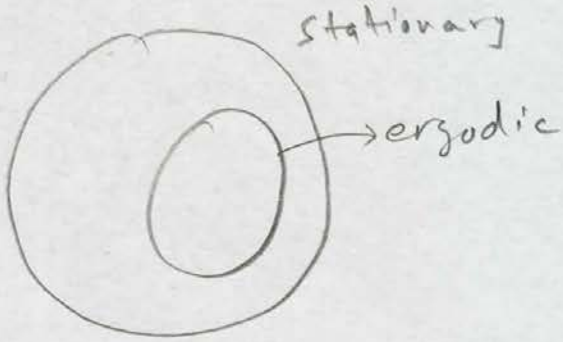
= One for All and All for one".

→ In Ex.2 process ~~one~~ each single realization manifests (given sufficiently long time) all the stats of the whole ensemble.

or All the stats shows itself in each single (long-enough) realization

→ That is not the case in Ex.1 process.

→ In fact, within stationary processes, we can define a further subclass of "Ergodic" processes, such that:



Defn: → "General Ergodicity"

→ An r.p. which is stationary and for which

$$\text{ensemble stats} = \text{Time stats}$$

is called "Ergodic".

→ Process in EX.2 is ergodic, whereas process in EX.1 is not.

Defn. As with stationarity, showing full ergodicity is hard (and often not of use), instead we often talk about < Ergodic in the mean, Ergodic in covariance } second order Ergodicity

Defn. "Ergodic in the mean"

$$\text{Ensemble Mean} = \text{Time Mean}$$

(here equality can be in a.s. or m.s. sense convergence)

Defn: "Ergodic in Covariance"

Ensemble covariance = Time covariance (here equality can be in a.s or m.s. sense ↓ convergence)

(3) Population vs. The individual

→ Ergodicity interpretation:

In general phenomena where entire populations statistical behaviour is same as the individual's are "ergodic"

→ e.g

Population	Individual (over time)
– Throw thousand coins	– Throw single coin thousand times
– Deduce frequency of H and T	– Deduce frequency of H and T

└── same ──┘

– Ergodic!

→ Second example:

Population	Individual (over time)
– Observe thousands of families in Islamabad to see which picnic spot is most liked	– observe a single family in Islamabad over several years to see which is their favourite picnic spot

→ Not necessarily same result

[e.g. may depend on the single family you choose to observe]

– In fact, human behaviour is, in general, highly non-ergodic.

4) Some Helpful Laws

ⓐ → Weak Law of Large Numbers (WLLN)

→ Let $X_1, X_2, \dots X_n \dots$ be i.i.d RVs with finite (and constant) mean $E[X_k] = \mu$ and finite variance (i.e. μ is the ensemble mean). Let the time-mean be denoted as

$$\bar{X}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

→ Then we have

$$\lim_{n\to\infty} P\left[|\bar{X}_n - \mu| \geq \epsilon\right] = 0 \quad \text{for any } \epsilon > 0$$

→ i.e the ~~ens~~ time-mean asymptotically approaches the ensemble ("true") mean in probability.

→ Such statement is often written briefly as

$$\bar{X}_n \xrightarrow{P} \mu \qquad \left(\begin{array}{c}\text{convergence}\\ \text{in probability}\end{array}\right)$$

(b)

→ Strong Law of Large Numbers (SLLN)

→ Let $X_1, X_2, \ldots X_k$ be iid RVs with finite (and fixed) ~~ensemb~~ mean $E[X_k] = \mu$ and with $E[X_k^2] < \infty$, then

$$\bar{X}_n \xrightarrow{a.s.} \mu \quad \text{and} \quad \bar{X}_n \xrightarrow{m.s.} \mu \qquad \left(\begin{array}{c}\text{"almost sure convergence"}\\ \text{and}\\ \text{"Mean square" convergence}\end{array}\right)$$

where $\bar{X}_n \triangleq \frac{1}{n}\sum_{k=1}^{n} X_k$ is the time mean.

Note: $\bar{X}_n \xrightarrow{a.s} \mu \overset{\text{expanded}}{\Rightarrow} P\left[\lim_{n\to\infty} |\bar{X}_n - \mu| < \epsilon\right] = 1$ for any $\epsilon > 0$

$$\bar{X}_n \xrightarrow{m.s} \mu \Rightarrow \lim_{n\to\infty} E\left[(\bar{X}_n - \mu)^2\right] = 0$$

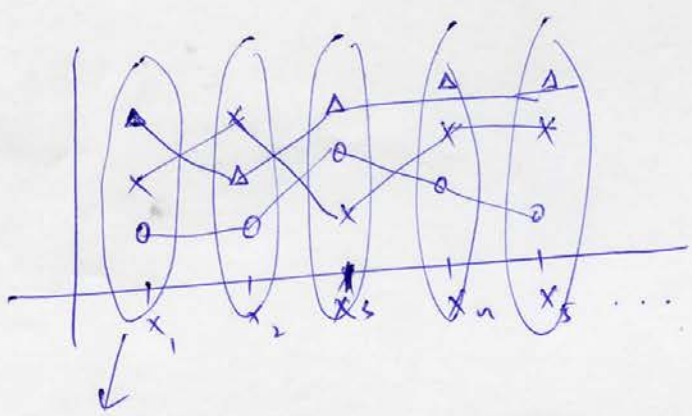
(5) Ergodicity — some formal Definitions.

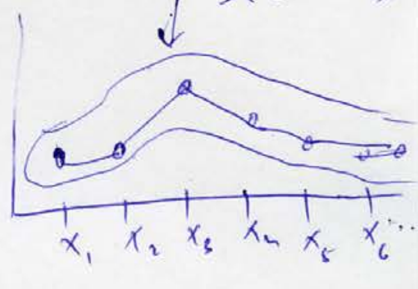
(a) "Ensemble" : collection of realizations of r.p.

(b) "Ensemble Statistics" : Statistics obtained from all possible realizations of an r.p. (= "true statistics")

(c) "Time Statistics" : statistics obtained from (infinitely) long single realization of an r.p.

e.g. $[X_1 \; X_2 \; X_3 \; X_4 \; X_5 \cdots]$ with $F_{X_1}, F_{X_2}, F_{X_3}, F_{X_4}, F_{X_5}$





→ ensemble stats of X_1 or r.p. X_t at $t = 1$ (etc.)

(d) "Ergodic" : An r.p. is ergodic if its ensemble stat (= "true stat") is the same as its time-stat

→ clearly, for that to be even possible, process must be stationary (i.e $F_{X_1} = F_{X_2} = F_{X_3} = \cdots$)

(e) "Strictly / completely" Ergodic : All ensemble stats = All Time stats (here equality can be in a.s. or m.s. convergence sense)

— a check for that :

— an r.p. is completely ergodic if any ensemble expectation

$$E\left[g(X_{t_1}, \ldots, X_{t_p}) \right]$$

can be "consistently" estimated from the corresponding time-average of a single-realization.

$$\frac{1}{n} \sum_{t=1}^{n} g\left(x(t+t_1), \ldots, x(t+t_p) \right)$$

where $g(\cdot)$ is an arbitrary function of a finite number of $X(t)$-variable.

— what we mean by "consistent" estimate, is covered later.

(f) "Linearly Ergodic" (or "Ergodic in the Mean") : A WSS process X_t with ensemble mean m is "ergodic in the mean" if

$$\lim_{n \to \infty} E\left\{ (\hat{m}_n - m)^2 \right\} = 0$$

where $\hat{m}_n = \frac{1}{n}\sum_{t=1}^{n} x_t$ is the time-average of the r.p.

→ How to check?

(1) Sufficient condition

uncorrelated will give new info about sample space ↓

$$\begin{cases} \lim_{\tau \to \infty} r_x(\tau) = 0 \\ r_x(0) < \infty \end{cases}$$

(i.e. r.v.'s should be asymptotically uncorrelated)

(2) Necessary & sufficient condition

$$\lim_{n \to \infty} \frac{1}{n} \sum_{\tau=0}^{n-1} r_x(\tau) = 0$$

(which clearly holds if $\sum_{\tau=0}^{\infty} r_x(\tau) < \infty$)

(g) = Second-order Ergodic : A WSS r.p. with ensemble covariance $r(\tau)$ is "second-order Ergodic" if

$$\lim_{n \to \infty} E\left\{\left(\hat{r}_n(\tau) - r(\tau)\right)^2\right\} = 0$$

where $\hat{r}_n(\tau) = \frac{1}{n}\sum_{t=1}^{n-\tau}(x_t - m)(x_{t+\tau} - m) \quad \tau \geq 0$

is the time-covariance of the r.p.

→ note that if m is unknown, one may replace it by the time mean $\hat{m}_n = \frac{1}{n}\sum_{t=1}^{n} x_t$.

→ Conditions for second-order ergodicity a bit more involved (and process dependent)

→ see e.g. Th. 2.5 and Th. 2.6 in textbook-1 [Lindgren].

6 Estimation

→ A large variety of applications require the estimation of mean and covariance of a stoch. process from a measured realization.

(a) → How can we compare different estimators?
→ What should the properties of a good estimator be?

→ e.g. true value θ, our estimate of it $\hat{\theta}^*_n$ ← estimate; n ← amount of data used.

① Estimator should be "unbiased":

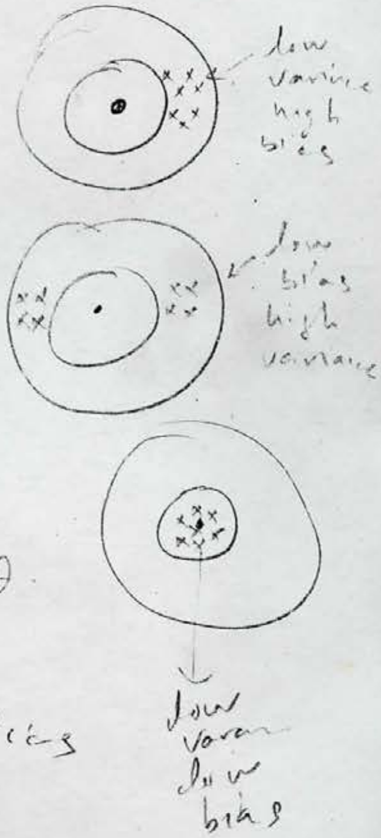
$$B \triangleq \theta - E\{\hat{\theta}^*_n\}$$ ← on avg. how far is $\hat{\theta}^*_n$ from θ. We want $B=0$, i.e. $E\{\hat{\theta}^*_n\} = \theta$

② Or, at least "asymptotically unbiased"

$$\lim_{n\to\infty} E\{\hat{\theta}^*_n\} = \theta$$

③ Should be "consistent"

$$\lim_{n\to\infty} \mathrm{Var}\{\hat{\theta}^*_n\} = 0$$



- True value is bullseye (θ).
- Estimator is a person throwing darts to hit θ.
- Each dart throw is an estimate ($\hat{\theta}_n$)
- First thing for a good marksman is lack of bias
- But that alone isn't enough!
- We also need low variance.
- We hit bullseye most often if both bias and variance are low.

(b)

— Estimating the mean of an ergodic process (from one realization)

$$\hat{m}_n^* = \frac{1}{n}\sum_{t=1}^{n} x_t$$

(1) check bias.

$$B = m_x - E\{\hat{m}_n^*\} = m_x - E\left\{\frac{1}{n}\sum_{t=1}^{n} x_t\right\}$$

$$= m_x - \frac{1}{n}\underbrace{\sum_{t=1}^{n} E\{x_t\}}_{n m_x} = m_x - m_x = 0$$

(2) check consistency.

— It can be shown that (see th. 2.4)

(1) if $\sum_{t=0}^{\infty} r(t)$ is convergent.

(2) then $V[\hat{m}_n^*] \simeq \frac{1}{n}\sum_t r(t)$

(3) then

$$\lim_{n\to\infty} V[\hat{m}_n^*] = 0$$

Thus $\hat{m}_n^*$ is conditionally consistent (depends on cov. fn.)

— i.e consistent! ⟶

— another way of saying the above is that

~~as $n \to \infty$~~

$$\lim_{n\to\infty} E\{(\hat{m}_n^* - m_x)^2\} \to 0$$ (converges in the mean sq. sense)

and

$$\lim_{n\to\infty} P\{|\hat{m}_n^* - m_x| > \epsilon\} \to 0 \quad \forall \epsilon > 0$$

(converges with probability one)

ⓒ Estimating the covariance

— Two candidates

can replace by $\hat{m}^*_n$

$$\hat{r}_n(\tau) = \frac{1}{n-\tau}\sum_{t=1}^{n-\tau}\left(x(t)-m_x\right)\left(x(t+\tau)-m_x\right)$$

or

$$r^*_n(\tau) = \frac{1}{n}\sum_{t=1}^{n-\tau}\left(x(t)-m_x\right)\left(x(t+\tau)-m_x\right)$$

— check biases

$$E\left\{\hat{r}_n(\tau)\right\} = \frac{1}{n-\tau}\sum_{t=1}^{n-\tau}\underbrace{E\left\{\left(x(t)-m_x\right)\left(x(t+\tau)-m_x\right)\right\}}_{r(\tau)}$$

$$= \frac{1}{n-\tau}(n-\tau)\,r(\tau) = r(\tau) \quad \leftarrow \text{unbiased.}$$

$$E\left\{r^*_n(\tau)\right\} = \frac{1}{n}\sum_{t=1}^{n-\tau} r(\tau) = \frac{n-\tau}{n}\,r(\tau) \leftarrow \text{biased}$$

$\rightarrow \frac{n}{n}$ as n grows large

$$\lim_{n\to\infty} E\left\{r^*_n(\tau)\right\} = \lim_{n\to\infty}\frac{n-\tau}{n}\,r(\tau) = r(\tau) \leftarrow \text{asymp. unbiased}$$

— $\hat{r}_n$ might seem better, however, we always use r^*_n

— see real reason in chapter 4

— It can be shown that

$$E\left\{\left[r^*_n(\tau)-r(\tau)\right]^2\right\} \le E\left\{\left[\hat{r}_n(\tau)-r(\tau)\right]^2\right\}$$

— Also notice that due to presence of τ in denominator of $\hat{r}_n(\tau)$ the variance will vary with τ (not good).

- Finally, r_n^* is consistent for a stationary Gaussian process ~~with~~ that has converged cvf (i.e $\sum_{\tau=0}^{\infty} r(\tau) < \infty$)

(th. 2.6) ← proof.

(3) Examples & discussions.

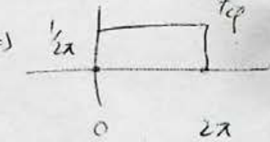
(3.1) Frequencies can be estimated from $r(\tau)$

→ As shown in Th 2.3 (Pls go through the proof carefully)

for $X(t) = A\cos(2\pi f_0 t + \varphi)$, A and φ independent r.v's

↙ random harmonic oscillator

f_A ↙ $U[0, 2\pi]$ ↙ $\Rightarrow$



we have

$$E[X(t)] = 0$$

$$V[X(t)] = \frac{1}{2}E[A^2]$$

$$r(\tau) = \sigma^2 \cos 2\pi f_0 \tau \quad \text{(note cvf has same frequency as } x(t))$$

→ similarly for superposition of random harmonic oscillations

$$X(t) = A_0 + \sum_{k=1}^{n} A_k \cos(2\pi f_k t + \varphi_k)$$

$A_0, \ldots A_n$ and $\varphi_1 \ldots \varphi_n$ be indep. $\varphi_k \sim U[0, 2\pi]$

we have

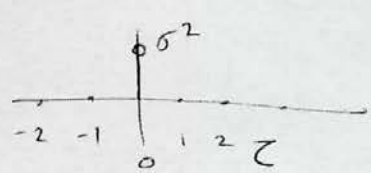
$$r(\tau) = E[A_0^2] + \sum_{k=1}^{n} \frac{1}{2}E[A_k^2]\cos(2\pi f_k \tau)$$

↙ Thus, $r(\tau)$ contains info on f_k's. → If we can estimate $r(\tau)$ from realization, we can find f_k's (using spectral estimation methods ← Later!)

(3.2) Estimating $V[m_n^*]$ from covariance function (or, how good is my mean est?)

(a) → For uncorrelated data X_t

i.e $r(\tau) = \begin{cases} \sigma^2 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases}$



→ we can estimate $V[m_n^*]$ using th. 2.4

→ for that first check the condition that $\sum_{t=0}^{\infty} r(t)$ is convergent

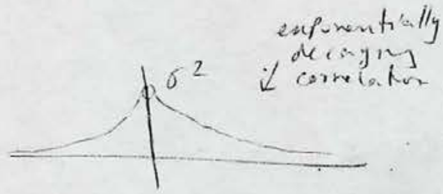
→ here $\sum_{t=\infty}^{\infty} r(t) = \sigma^2$ (convergent) → so by th. 2.4 $\hat{m}_n^*$ is consistent and further

$$V[\hat{m}_n^*] \approx \frac{1}{n} \sum_t r(t) = \frac{\sigma^2}{n}$$ decrease with increasing n.

$$D[\hat{m}_n^*] \approx \frac{\sigma}{\sqrt{n}}$$ (standard deviation)

(b) → For correlated data X_t with

$r_x(\tau) = \sigma^2 e^{-|\tau|}$



→ check: $\sum_{t=-\infty}^{\infty} r(t) = \sigma^2 \sum_{t=-\infty}^{\infty} e^{-|t|} = \sigma^2\left(1 + 2\sum_{t=1}^{\infty} e^{-|t|}\right) = \sigma^2\left(1 + 2\sum_{t=1}^{\infty}\left(\frac{1}{e}\right)^t\right)$

$$= \sigma^2\left(1 + 2\frac{1/e}{1 - 1/e}\right) = \sigma^2 \frac{e+1}{e-1}$$ → convergent!

using G.S

$\sum_{k=0}^{\infty} a r^k = \frac{a}{1-r}$ (for $|r| < 1$), we have $\sum_{t=1}^{\infty} \left(\frac{1}{e}\right)^t = \sum_{t=0}^{\infty}\left(\frac{1}{e}\right)^t - 1$

$$= \frac{1}{1 - 1/e} - 1 = \frac{1 - 1 + 1/e}{1 - 1/e}$$

→ so, $V[\hat{m}_n^*] \approx \frac{1}{n}\sum_t r(t) = \frac{\sigma^2}{n}\frac{e+1}{e-1}$

$D[\hat{m}_n^*] \approx \frac{\sigma}{\sqrt{n}}\sqrt{\frac{e+1}{e-1}} = \frac{\sigma \, 1.471}{\sqrt{n}}$ ($\hat{m}_n^*$ for correlated data has higher S.D.)

↓

so for more correlated data you need more samples to get accurate estimate. ←

(e.g. height of population from sample → what if all samples are from same family? city)

(3.3) If $\{X_t\}$ is Gaussian process with mean m_X ~~and~~, then $\hat{m}_n^* = \frac{1}{n}\sum_{t=1}^{n} X_t$ is also Gaussian (normal) with mean m_X and variance $V[\hat{m}_n^*]$ i.e.

$$\hat{m}_n^* \sim \underset{\text{normal}}{\mathcal{N}}(\overset{\text{true mean}}{m_X}, \overset{\text{var.}}{V[\hat{m}_n^*]})$$

∴ → consequence of the fact that sum of Gaussians is also Gaussian (ch. 5)

(3.4) Confidence Interval (should know from basic probability)

→ suppose I give you an estimate of the mean as $\hat{m}_n^* = 0.4$

→ since it's an "estimate", it would be nice if I could say something about how sure I am about my estimate

→ e.g. I am 95% sure that the true mean is in the range of $[0.37, 0.43]$ (i.e. 0.4 ± 0.03) ($P(0.37 \le m \le 0.43) = 0.95$)

→ For Gaussian case, we can give this interval by using the standard Gaussian quantiles. E.g., for confidence level $(1-\alpha)$% we have

$P(-\lambda_{\alpha/2} \le Y \le \lambda_{\alpha/2}) = 1-\alpha$, $Y \sim N(0,1)$

$$P\left(\hat{m}_n^* - \lambda_{\alpha/2} D[\hat{m}_n^*] \le \hat{m} \le \hat{m}_n^* + \lambda_{\alpha/2} D[\hat{m}_n^*]\right) = 1-\alpha$$

(quantile)

with $I_m : \left\{\hat{m}_n^* - \lambda_{\alpha/2} D[\hat{m}_n^*],\ \hat{m}_n^* + \lambda_{\alpha/2} D[\hat{m}_n^*]\right\}$

→ The above I_m also serves as a good guess for non-Gaussian case.

→ Pls see Ex. 2.16 in the book.

Lecture 8

① A Tale of Two Processes < Poisson (CTDS) / Gaussian (CTCS)

— Both extremely important and in common use.

— Poisson: often naturally occurs where we keep a "count" of occurances (arrival, birth, calls ...)

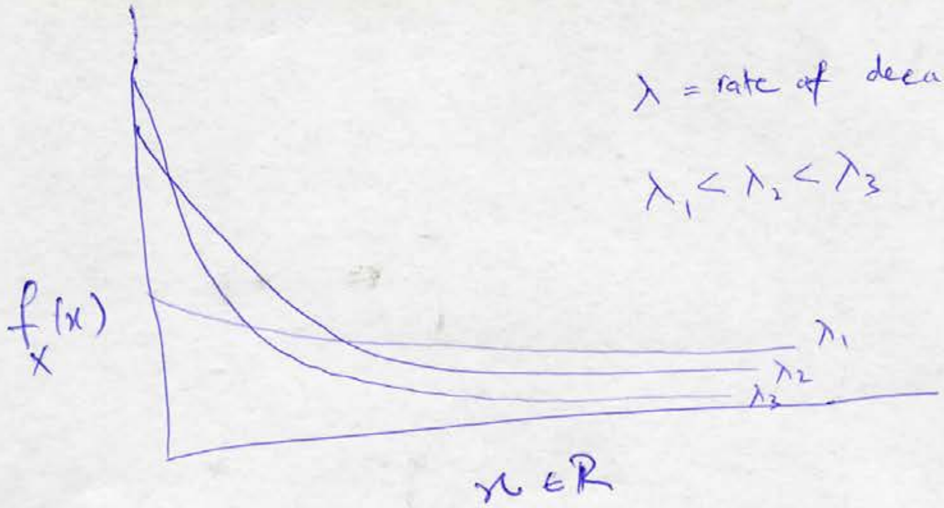
— Gaussian or "Normal": A lot of random phenomena are in fact cumulative effects of several underlying random processes. Such "cumulative effects" are often best represented by the Gaussian/normal distribution (a fact that is also mathematically embodied by the "Central Limit Theorem")
↓
later!

② Some quick Background.

ⓐ Exponentially Distributed RV (continuous)

— suppose you have a random variable which is non-negative and whose relative likelihood of taking larger numbers drops exponentially

— such RV is often best represented by the "Exponential Distribution".



→ probability of seeing larger values drops exponentially.

– Mathematically:

$$f_X(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad \boxed{\lambda > 0}$$

– notation

$$X \sim EXP(\lambda)$$

– moments:

$$E[X] = \frac{1}{\lambda}, \quad Var[X] = \frac{1}{\lambda^2}$$

(see, e.g., Kobayashi)

(b) Poisson and the Law of Rare Events

— Consider a situation where a large number of events can occur (independently of each other) but each is very rare (i.e. only a few actually occur)

- e.g., almost the whole world can call ~~by~~ my number, but on a given day only a (relatively) "few" people actually do!

- DT RVs that ~~keep~~ represent ~~"actually~~ "how many of these independent rare events actually occur in a given interval" are often best represented by the Poisson Distribution → (with some additional conditions)

- Mathematically

$$P\left(k \text{ events occur in interval } t\right) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \quad (k = 0, 1, 2, \ldots)$$

where $\lambda > 0$ is the "intensity" or "number of events per unit time" (fixed).

Denoted $\text{Pois}(\lambda t)$

or

$$p_X(k; \lambda t) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

$\lambda > 0$
$k = 0, 1, 2, 3, \ldots$

$E[X] = \lambda t$
$V[X] = \lambda t$

- Where does it come from?
  → Limiting case of Binomial Distribution!

- Binomial? (Recall:)

  - If there are (n) events that can occur independently of each other such that each one either occurs with probability (p) or does not occur with probability $(1-p)$, then the probability that (k) of them will occur follows the "Binomial Distribution" given as

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

$(1-p)^{n-k}$ ← n-k do not occur.

p^k → k independent event occurs each with probability p i.e. $\underbrace{p \cdot p \cdot p \cdot p \cdot p \cdots p}_{k}$

$\binom{n}{k}$ ↓ # of ways k events can occur out of n

- what happens as $n \to \infty$? (i.e. total number of events that can occur becomes extremely large?)

$$\lim_{n \to \infty} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k e^{-\lambda}}{k!}$$

↑ if the limit exists

(where $\lambda = np$ or $p = \lambda/n$)

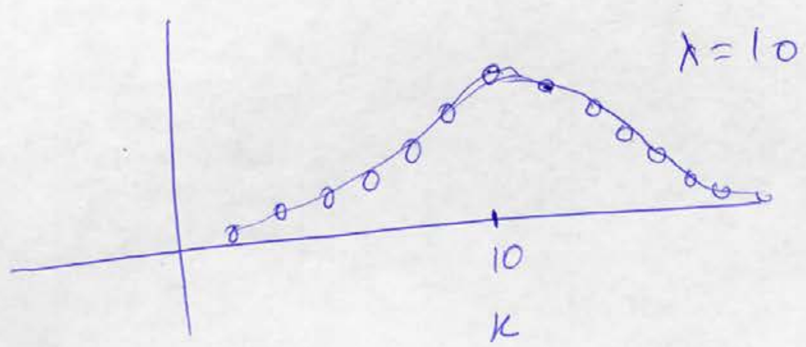
λ ↓ expected number of events to occur

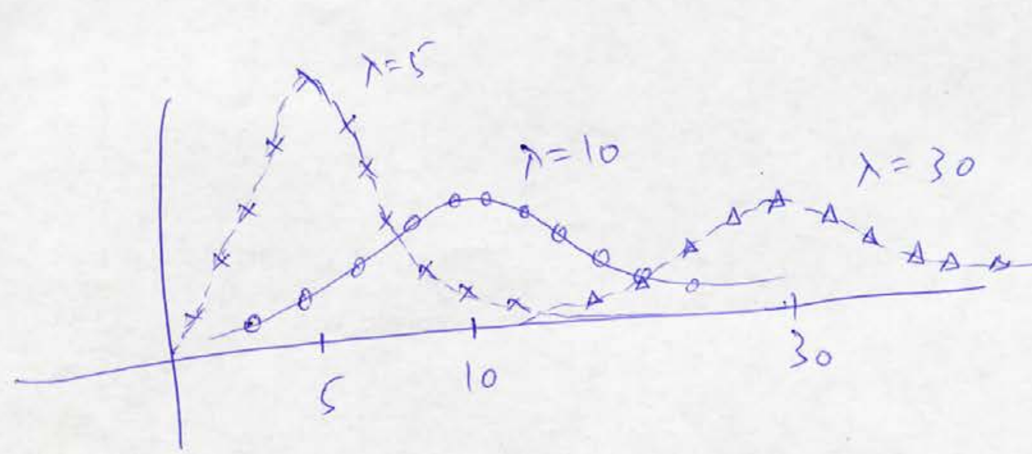
— So how does Poisson distribution look?

→ for unit time interval (i.e $t = 1$) we may write the Poisson Distribution as

$$p(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

→ now for $\lambda = 10$ we get

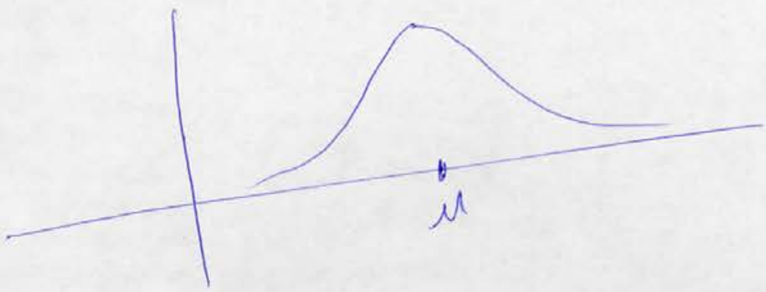




→ i.e expected value per unit time (λ) has highest probability of occuring here, while values larger or smaller than λ have fast dropping probabilities.

(c) Gaussian Distribution & The Central Limit theorem (CLT)

(i) In a large number of physical phenomena, things tend to be more clustered towards a "mean", and increasing deviations from it are increasingly less likely.



(ii) Many random phenomena are in fact an overall (or "average") effect of several underlying random occurances.

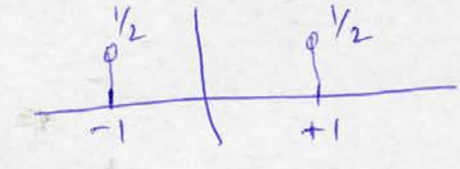
- CLT tells us that random phenomena thus created often follow a special distribution called the Gaussian/Normal distribution (which also exhibits the property described in (i))

- e.g. consider a coin-tossing game where you win 1 rupee for a Head and loose 1 rupee for a tail.

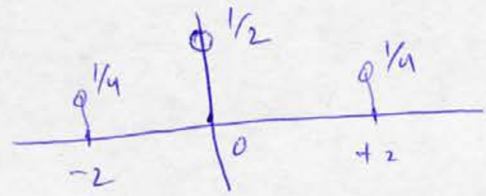
- X = overall amount you win or loose

- let's look at the distribution of X

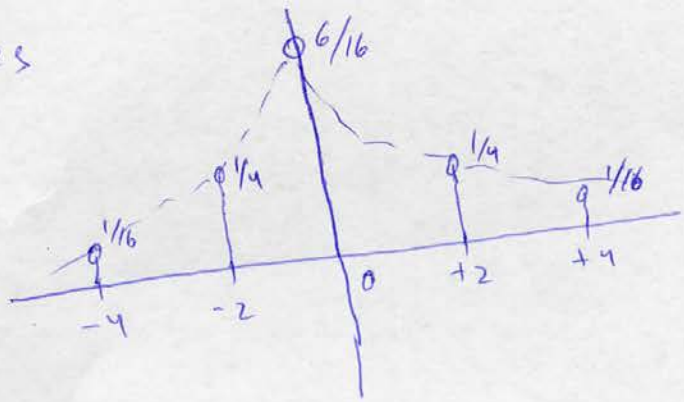
- for one toss



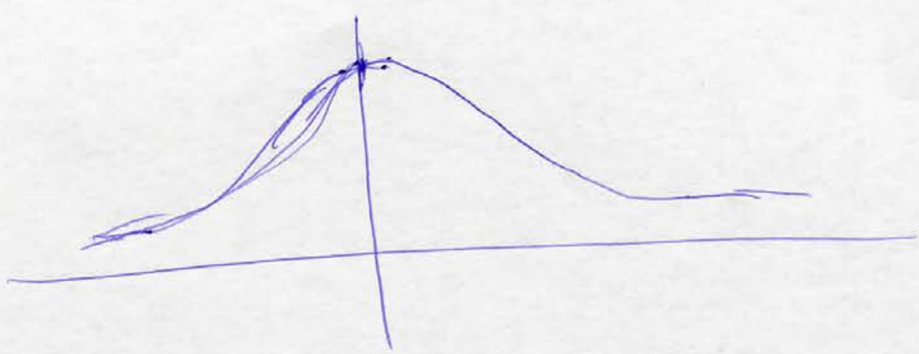
- for two tosses



- for four tosses



- easy to imagine shape as no. of tosses increases



- In fact CLT tells us that this shape is described by the formula that we call 'Gaussian/normal' Distribution.

CLT
— In words:

Basic CLT: The average (behaviour) of i.i.d r.v.s follows a normal distribution

Generalized CLTs: under some rather loose constraints, average (behavior) of r.v.s that are not independent and/or not identically distributed also follows the normal distributions

e.g Basic CLT: if we have i.i.d RVs. $\{X_k\}$ then, their "average" defined as

$$Z_n = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$$

Follows a Gaussian distribution in the limiting case $n \to \infty$.

i.e $$Z_n \xrightarrow[n\to\infty]{D} G.D.$$

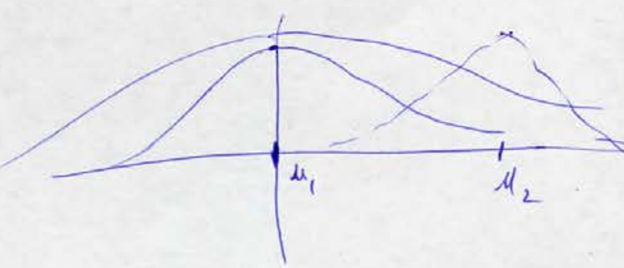
— Mathematically we write a G.D. as

$$f_X(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

→ where μ & σ^2 represent the mean and variance of the Gaussian r.v. X.

— notation:

$$X \sim \mathcal{N}(\mu, \sigma^2)$$



— Additional important property:

— G.D. is full characterized by mean & variance.

— All higher moments can be computed if μ & σ^2 known

— $$E\left[(X-\mu)^k\right] = \begin{cases} 0 & \boxed{k \text{ odd}} \\ \sigma^k (k-1)(k-3)\dots(1) & \boxed{k \text{ even}} \end{cases}$$

— Link: Poisson & Gaussian

— In some ways, Poisson and Gaussian are DS and CS counterparts

— In fact, for large λ values, Gaussian Distribution provides a good CS approximations of Poisson.

(3) Poisson Process

- we have previously seen the Poisson Process as a special case of renewal processes (which in turn are special case of point/counting processes).

- Let's define it formally now:

(a) Definition 1

- Consider a random process where point events occur at random times such that the spacings between the times $\{X_k\}$ defined as $X_k = t_k - t_{k-1}$ are i.i.d and follow the ~~Poisson~~ exponential distribution, i.e.,

$$\{X_k\} \sim \text{i.i.d} \quad \text{and } f_{X_k}(x) = f_X(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

(where $\lambda > 0$ is the rate of decay)

- Then the process that maintains a count of events in time $[0,t]$, say $N(t)$, is a Poisson Process and follows the Possion distribution, i.e

$$N(t) \sim \text{Pois}(k, \lambda^{\circ}) \quad \text{when } \lambda^{\circ} = 1/\lambda$$

i.e. $$P(\underset{\text{in interval } h}{N(t) = k}) = \frac{(\lambda^{\circ} h)^k e^{-\lambda^{\circ} h}}{k!}$$

(b) Definition 2

Consider a stochastic process $\{X(t), t \in T\}$ that has

(i) Independent increments

i.e. $X(t_2) - X(t_1)$
$X(t_3) - X(t_2)$
$\vdots$
$X(t_n) - X(t_{n-1})$

are independent for every selections
$t_1 \le t_2 \le \ldots \le t_n \in T$

i.e Number of events occurring in non-overlapping intervals are independent of each other

(ii) Stationary increments

i.e Distribution of $X(t+h) - X(t)$ (# of events in interval h) depends only on h (length of the interval) and not on t.

(iii) Simply-Increasing Property

i.e. — It is non-decreasing with integer jumps
and The probability of more than one event in an infinitesimaly small interval approaches zero (*)
(i.e. "two events do not occur at exact same time")

(*) faster than h

$$\Rightarrow \frac{P[X(t+h) - X(t) > 1]}{h} \to 0 \text{ as } h \downarrow 0$$

(iv) $X(0) = 0$ (i.e count starts at zero, no previous accumulation)

Such a process will in fact follow a Poisson Distribution and is called a "Poisson Process".

<u>Mathematically</u> A poisson process with intensity (arrival rate) $\lambda > 0$ has

$$P\left[X(t+h) - X(t) = k\right] = e^{-\lambda h}\frac{(\lambda h)^k}{k!} \qquad k = 0, 1, 2, \ldots \quad h > 0$$

$$E\left[X(t)\right] = V\left[X(t)\right] = \lambda t$$

$$E\left[\frac{X(t+h) - X(t)}{h}\right] = \lambda$$

where $X(t) = \#$ of events in $[0, t]$

and $X(t+h) - X(t) = \#$ of events in an interval of length h

(which by "stationary increment" property is independent of t)

→ In addition, it can be shown that

— $C[X(s), X(t)] = r(s,t) = \lambda \min(s,t)$

— Poisson process is NOT WSS

— both conditions fail in fact

— $m(t) = E[X(t)] = \lambda t$ ← time dependant mean

— $r(s,t) = \lambda \min(s,t)$ ← does not depend on $|s-t|$ only.

(c) A parting note: It is important to keep in mind that while Poisson Process [as defined above] is a very commonly used arrival (counting) model, it is not the only one and does not, in particular, apply to situations where things may happen in clusters (e.g. student arrivals in lecture hall)..

— Extensions to Poisson required

or use other counting processes.

(4) Gaussian Process (Normal)

(a) Defn 1 — A stochastic process $\{X(t), t \in T\}$ is a Gaussian Process if every linear combination (i.e $a_1 X(t_1) + \dots + a_n X(t_n)$) of its values follows the Gaussian (normal) distribution.

→ note that this definition also implies that $X(t_k)$ for any fixed $t_k \in T$ also has a Gaussian distribution, i.e,

$$X(t_k) \sim \mathcal{N}\left(m(t_k), r(t_k, t_k)\right)$$

(since $X(t_k) = (0)(X(t_1)) + (0)(X(t_2)) + \dots + (1)(X(t_k)) + \dots (0)(X(t_n))$)

→ In fact, it can be shown that summation, differentiation, and integration of a Gaussian process always result in a value or a process that is also Gaussian (provided the said operations exist).

— So for instance, if $\{X(t)\,;\, t \in T\}$ is a Gaussian process then so are

Gaussian RP ← $Y_1(t) \triangleq X(t+h) + X(t) \quad (t, t+h \in T)$

Gaussian RV ← $Y_2 \triangleq \dfrac{X(t_1) + \dots + X(t_n)}{n}$

Gaussian RV ← $Y_3 \triangleq 0.5\,X(t_1) + 0.3\,X(t_2) + 0.4\,X(t_3)$

etc.

— Even limits of linear combinations (if they exist) will be Gaussian, e.g.

$$\lim_{h \to 0} \frac{X(t+h) - X(t)}{h} \longrightarrow \text{Gaussian RP if it exists.}$$

Imp: note To show that a process is Gaussian, it is not enough to show that each individual values of the RP is Gaussian.

i.e. → there can be a RP $\{X(t)\,;\, t \in T\}$ for which every $X(t_k) \sim \mathcal{N}$ $(t_k \in T)$ but still the process itself is not normal!!!

(b) <u>Defn 2</u> A stochastic process $\{X(t), t \in T\}$ is Gaussian if for every n and all $t_1 \dots t_n \in T$ the vector $[X(t_1) \dots X(t_n)]$ has a joint PDF represented by the n-dimensional Gaussian distribution.

— Recall: 1-D Gaussian distribution

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

$$\Rightarrow f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

— n-D Gaussian distribution (JPDF of n ~~Gaussian~~ RVs)

$$\underbrace{[X(t_1) \dots X(t_n)]}_{\triangleq \underline{X}} \sim \mathcal{N}(\underline{\mu}, \Sigma)$$

$$\Rightarrow f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})}$$

where $\underline{x} = \begin{bmatrix} x_{t_1} \\ x_{t_2} \\ \vdots \\ x_{t_n} \end{bmatrix}$; $\underline{\mu} = \begin{bmatrix} E(X_{t_1}) \\ \vdots \\ E(X_{t_n}) \end{bmatrix} \triangleq \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$

and

$$\Sigma = \begin{bmatrix} r(t_1,t_1) & r(t_1,t_2) & \cdots\cdots & r(t_1,t_n) \\ r(t_2,t_1) & \cdots & \cdots & \vdots \\ \vdots & & & \\ r(t_n,t_1) & \cdots & \cdots & r(t_n,t_n) \end{bmatrix}$$

where $r(t_i,t_k) \triangleq C\left[X(t_i), X(t_k)\right]$

→ Σ is called "covariance Matrix".

- <u>Imp. note</u> If the covariance matrix (Σ) is non-invertible then the vector is said to have a "singular" distribution and has no density function (PDF) in $\mathbb{R}^n$.

ⓒ Some interesting Properties of the Gaussian Process (and other remarks)

<u>(i)</u> A weakly Stationary G.P. is also strictly Stationary

→ This is the consequence of the fact that a G.P. is fully characterized by second-order statistics

— so, if its second-order stats exhibit stationarity then the ~~en~~ process is also strictly stationary (all of its higher-order stats will exhibit stationarity).

— e.g. Let $X(t)$ be WSS Gaussian process,

then $m_x(t) = m_x$

$r_x(t, t+\tau) = r_x(\tau)$

— now using definition of the Gaussian JPDF

$$f_{\underline{X}}(\underline{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})\Sigma^{-1}(\underline{x}-\underline{\mu})}$$

we see that $f_{\underline{X}}(\underline{x})$ depends only on τ (and not on t), since

$$\underline{\mu} = \begin{bmatrix} m_x \\ m_x \\ \vdots \\ m_x \end{bmatrix} \quad \text{and} \quad \Sigma = \begin{bmatrix} r(0) & r(1) & \cdots & r(n-1) \\ r(1) & r(0) & & \\ \vdots & & \ddots & r(0) \\ r(n-1) & \cdots & & \end{bmatrix}$$

$\underline{\mu}$: all same

Σ: since each covariance is independent of t.

— So the entire JPDF is "stationary"

$\Rightarrow$ strictly stationary.

(ii) Standard Gaussian RV.

→ we say an RV is "standard Gaussian" if it has mean 0 and variance 1, i.e

$$Y_t \sim \mathcal{N}(0,1) \quad Y_t \text{ is standard Gaussian RV.}$$

→ a non-standard Gaussian RV can be converted into standard Gaussian as follows. Let

$$X_t \sim \mathcal{N}(m, \sigma^2)$$

define $Y_t = \dfrac{X_t - m}{\sigma}$ then $Y_t \sim \mathcal{N}(0,1)$

and $\Pr\left\{X_t \le \alpha\right\} = \Pr\left\{Y_t \le \dfrac{\alpha - m}{\sigma}\right\} = \phi\left(\dfrac{\alpha - m}{\sigma}\right)$

↳ value can be read from CDF tables for standard Gaussians.

(5) Wiener Process (Brownian Motion)

— we previously saw the wiener process as the limit (if it exists) of the random walk

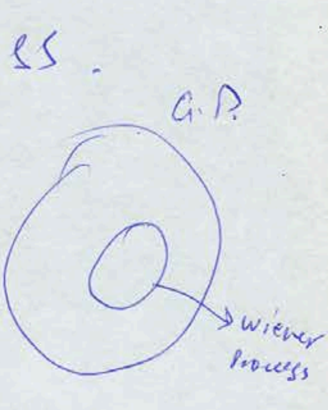
$$\lim_{\substack{n\to\infty,\ h\to 0 \\ \delta\to 0}} S_n = W(t) \quad \text{with } t = nh,\ \frac{\delta^2}{h} = \alpha$$

— and where $X_i = \begin{cases} +\delta & \text{with prob. } 1/2 \\ -\delta & \text{with prob. } 1/2 \end{cases}$

$\{X_i\}$ all i.i.d

(Recall, $S_n = X_1 + X_2 + \dots + X_n = S_{n-1} + X_n$)

— In fact, it can be shown the the Wiener process is a Gaussian Process.



i.e $W(t) \sim \mathcal{N}$

— Interestingly, its definition is rather similar to that of Poisson Process

<u>Defn</u> A Gaussian Process $\{X(t),\ 0 \leq t < \infty\}$ is a wiener process if

(i) $X(0) = 0$

(ii) the increments $X(t_2) - X(t_1),\ X(t_3) - X(t_2), \ldots$

$X(t_n) - X(t_{n-1})$ are independent for all $0 < t_1 < t_2 < \cdots < t_n$

(i.e non-overlapping intervals have independent increments)

(iii) The increment $X(t+h) - X(t)$ has a normal distribution $N(0, \sigma^2 h)$ for $0 \leq t \leq t+h$

→ note that from (iii) we conclude that

$$X(t) \sim N(0, \sigma^2 t)$$

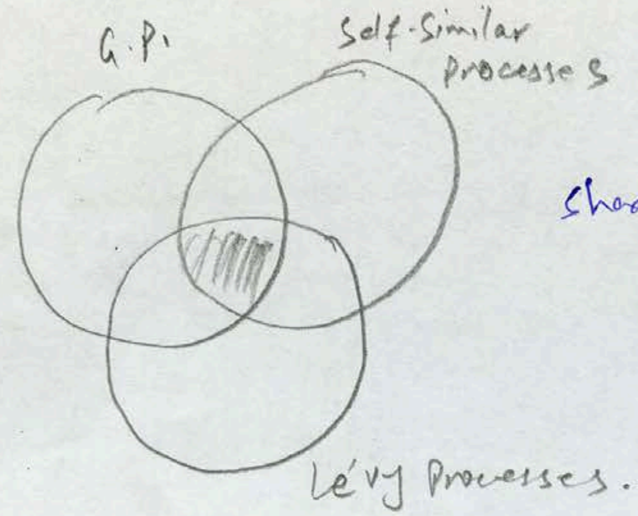
→ Covariance of a wiener process $X(t)$ is given by

$$C[X(s), X(t)] = r(s,t) = \sigma^2 \min(s,t)$$

— which implies the Wiener processes are non-stationary.

— In addition, a W.P. is everywhere non-differentiable
— It is very irregular no matter the magnification.
— It also belongs to two other classes of R.P.s

— Self similar processes
— Levy Processes



Shaded = Wiener Process (Standard Brownian Motion)

(6) Self-Similar Processes

— Simple words: If any zoomed-in (or zoomed out) version of the process exhibits the same statistics as the original.

— Mathematically: Stochastic process $\{X(t) ; 0 \leq t \leq \infty\}$ is called self-similar with index H if, for all n and $\lambda > 0$ and $t_1, \dots t_n$, the random vector

$$\left[\lambda^{-H} X(\lambda t_1) \ldots \lambda^{-H} X(\lambda t_n)\right]$$ ← Scaled version

has the same distribution as

$$\left[X(t_1) \ldots\ldots X(t_n)\right]$$ ← original version.

— It can be shown that a W.P. is Self-Similar if index $H = 1/2$

(7) Lévy Process

Defn A stochastic process $\{X(t),\ 0 \le t \le \infty\}$ is a Lévy process if

(i) $X(0) = 0$

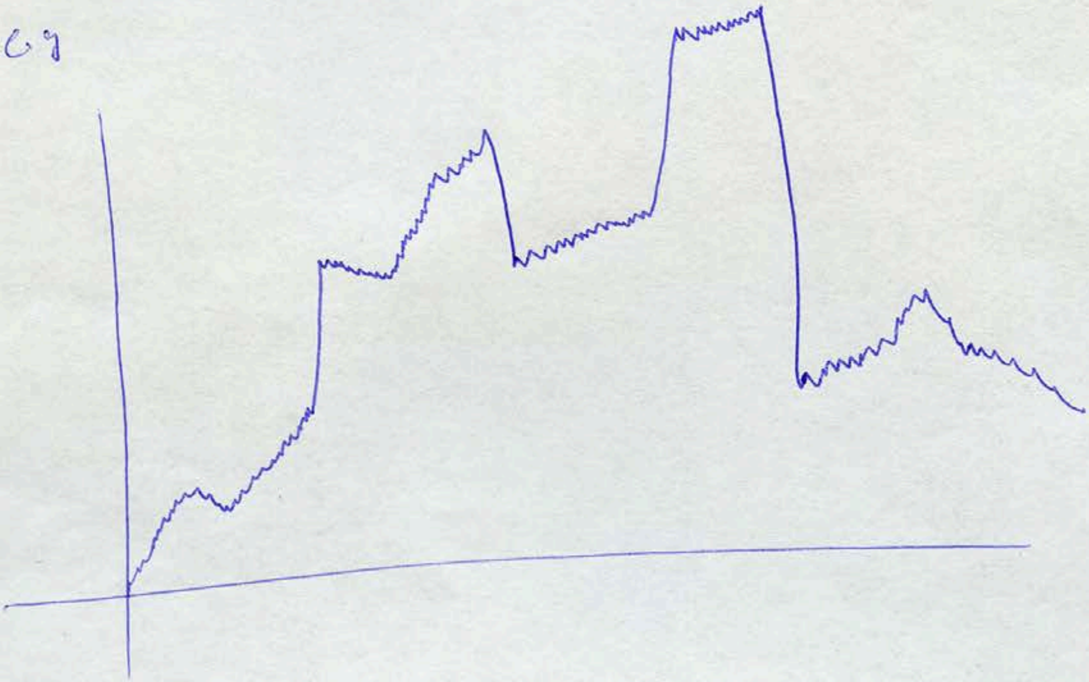
(ii) It has stationary increments

(iii) It has independent increments.

— Clearly, a W.P. satisfies all three conditions above (with the additional property that the increments are Gaussian) and is therefore a special Lévy Process.

Application: Lévy Processes are a large class of CTCS processes that are used to model random phenomena that exhibit periods of gradual continuous change combined with large and abrupt jumps.

e.g.



ES 544
Random Processes

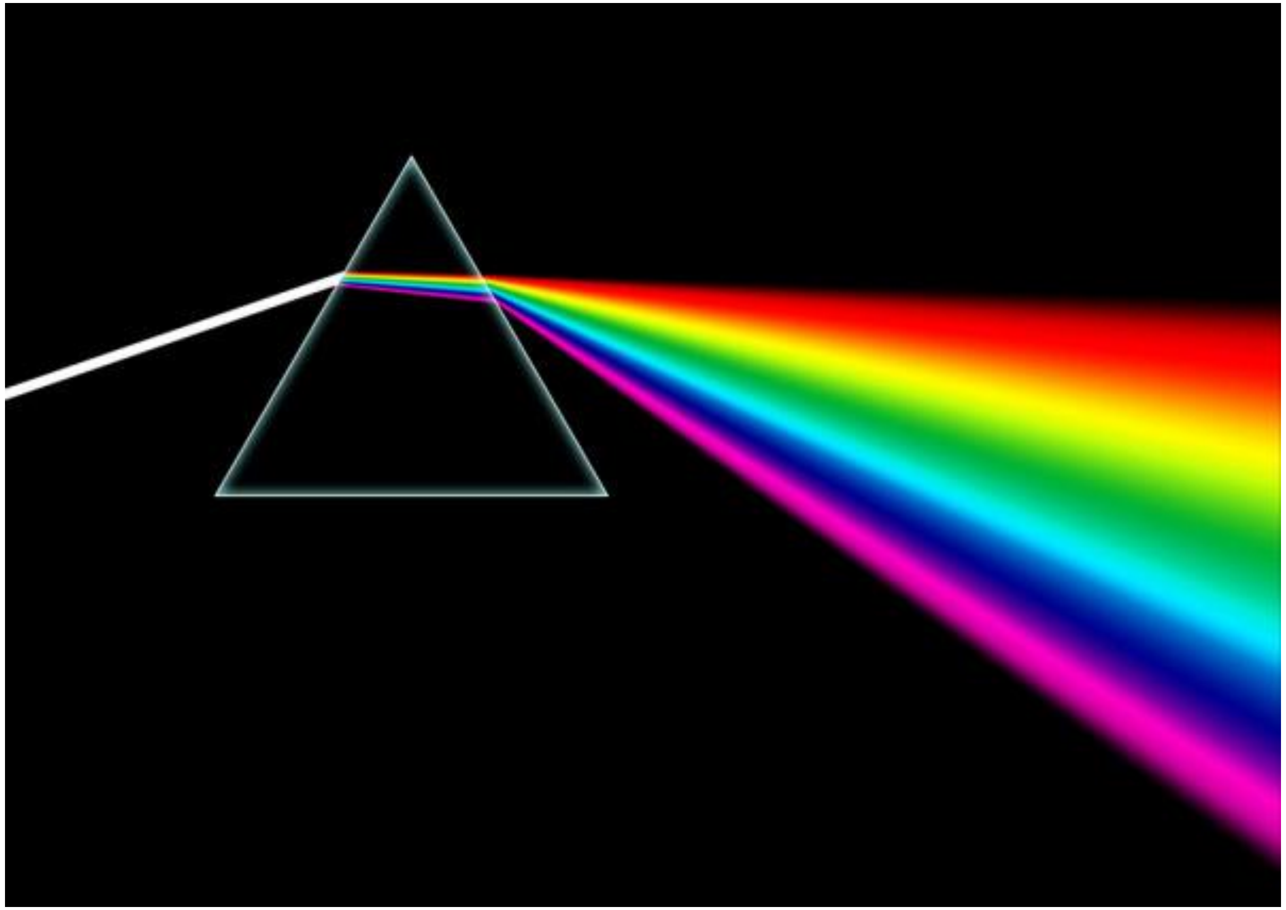
with

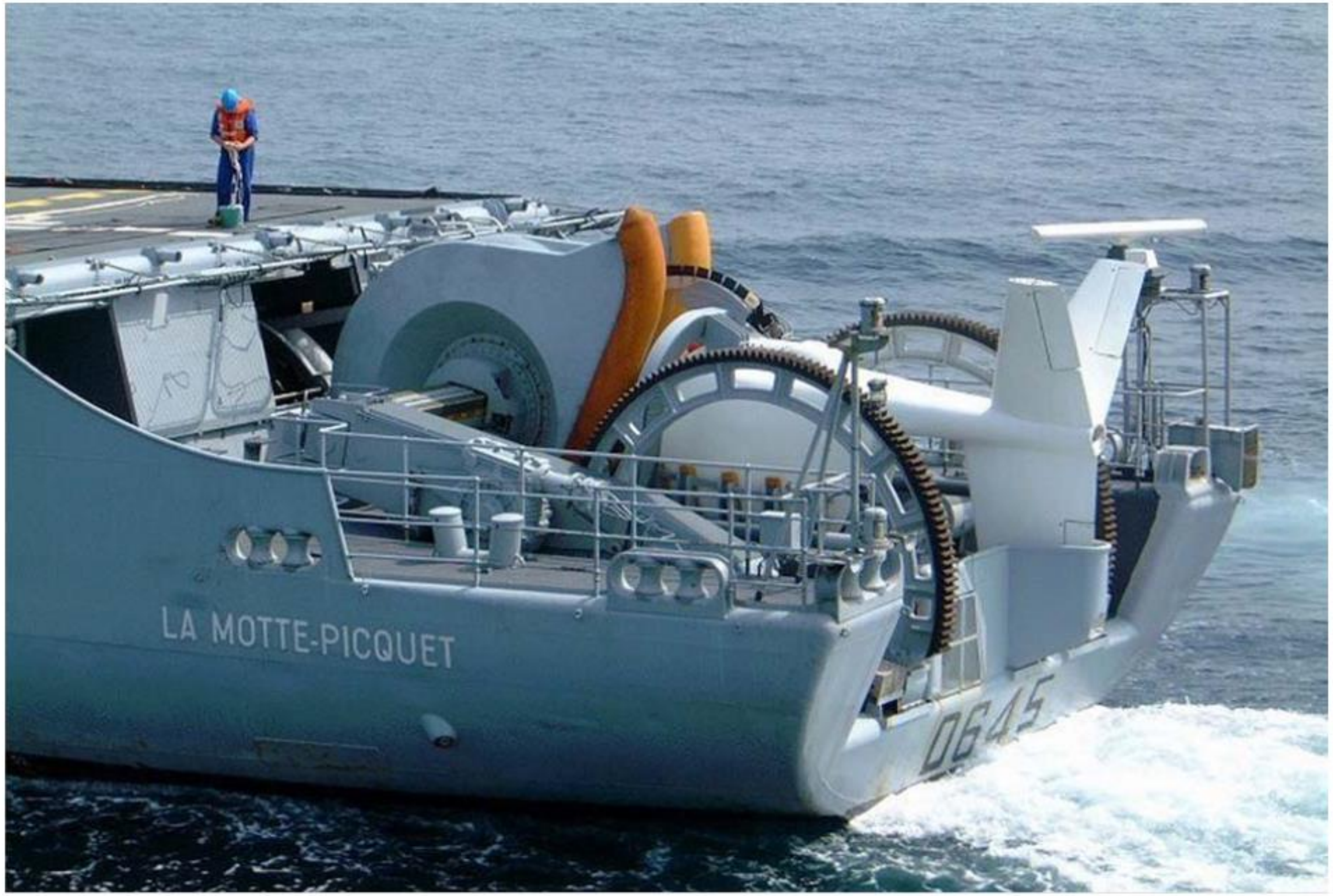
Dr. Naveed R. Butt

@

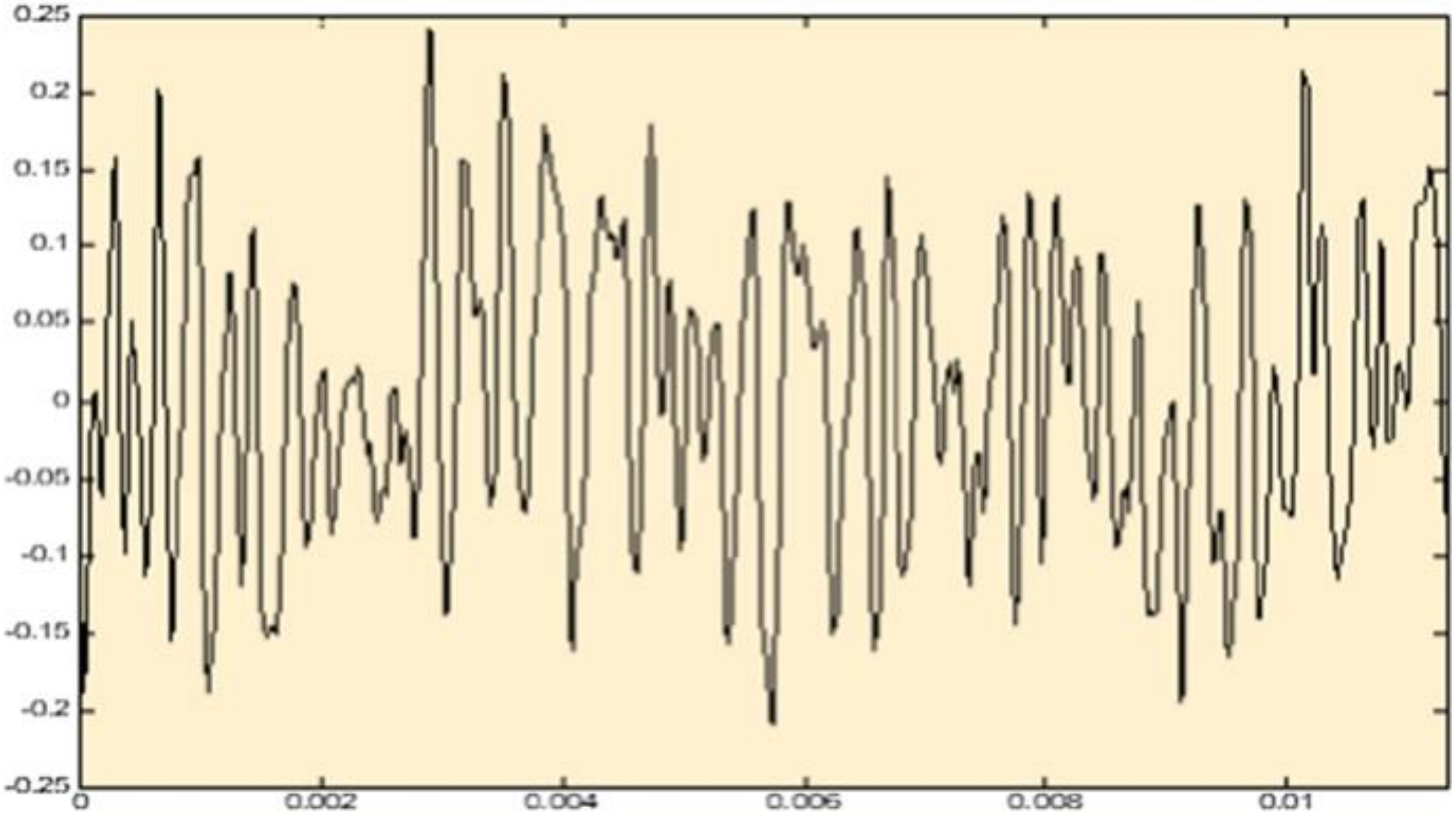
GIKI - FES

Spectra – *the Ghosts in Your Data*

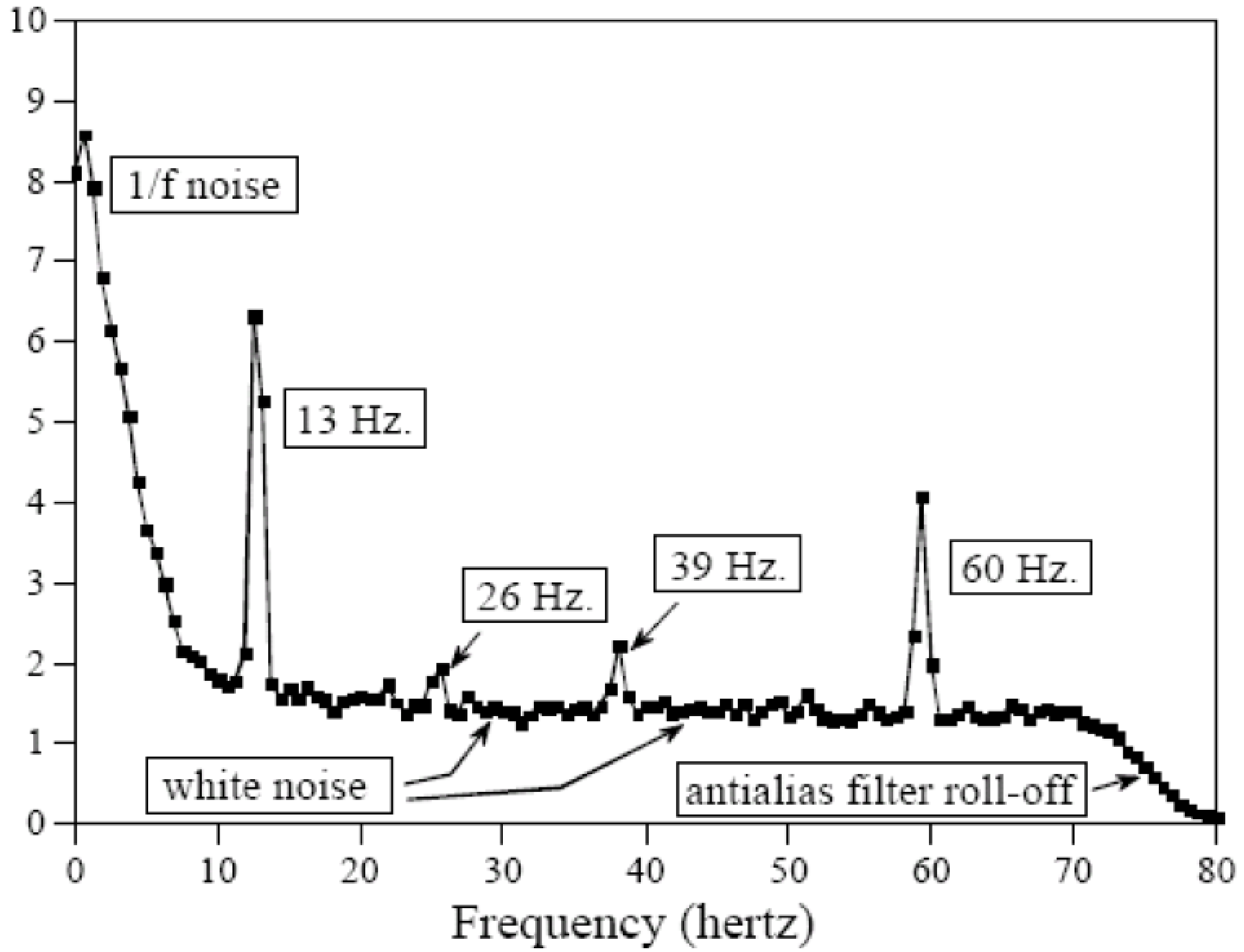




Sonar


0.25
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.2
-0.25
0
0.002
0.004
0.006
0.008
0.01

J. Boilly Del.
Geille Sculp.

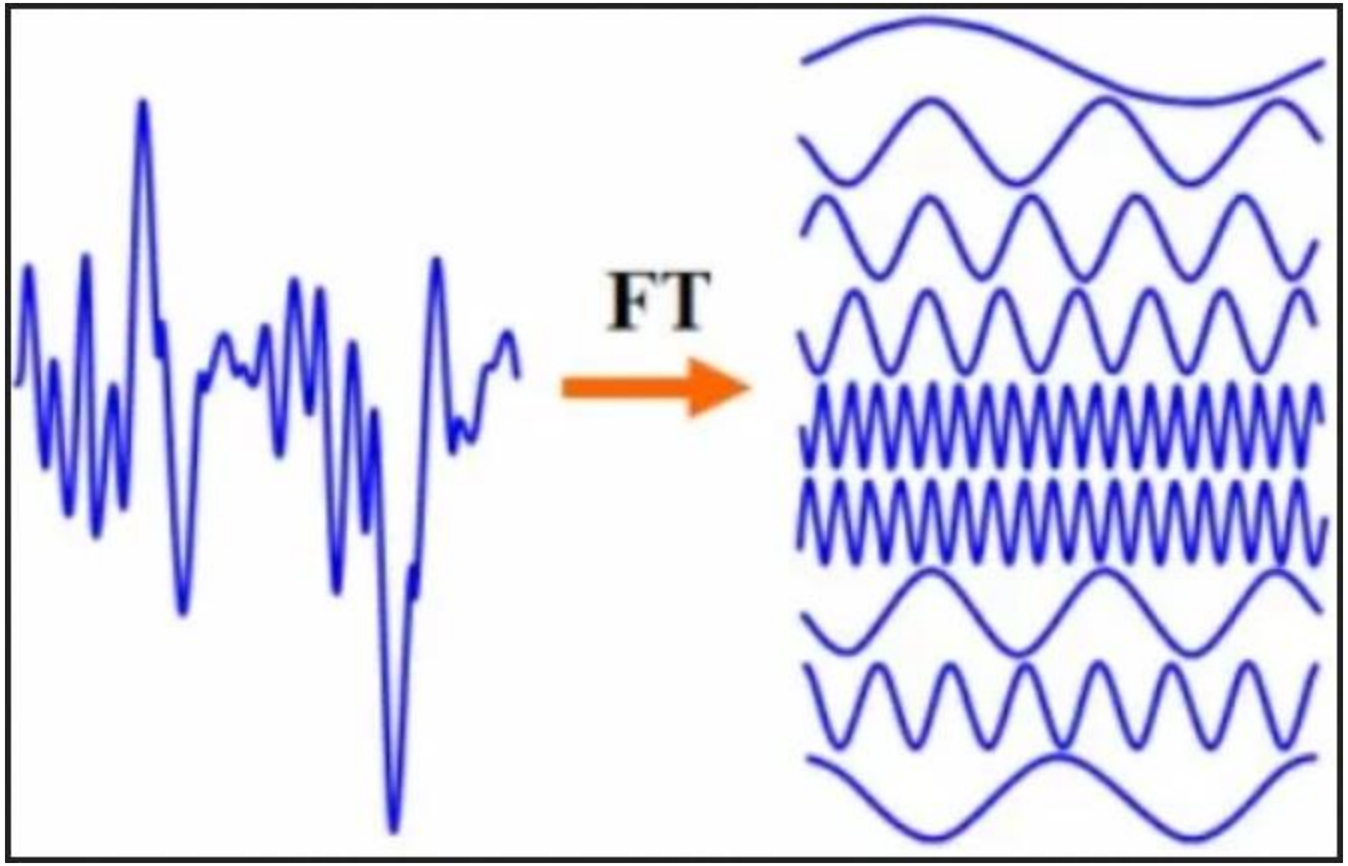

10
9
8
7
6
5
4
3
2
1
0
1/f noise
13 Hz.
26 Hz.
39 Hz.
60 Hz.
white noise
antialias filter roll-off
0
10
20
30
40
50
60
70
80
Frequency (hertz)

Q. Can we write processes/signals as sums of periodic functions (frequencies)?

This is exactly what Fourier Transform does – it tries to write every signal as a sum of sinusoids.

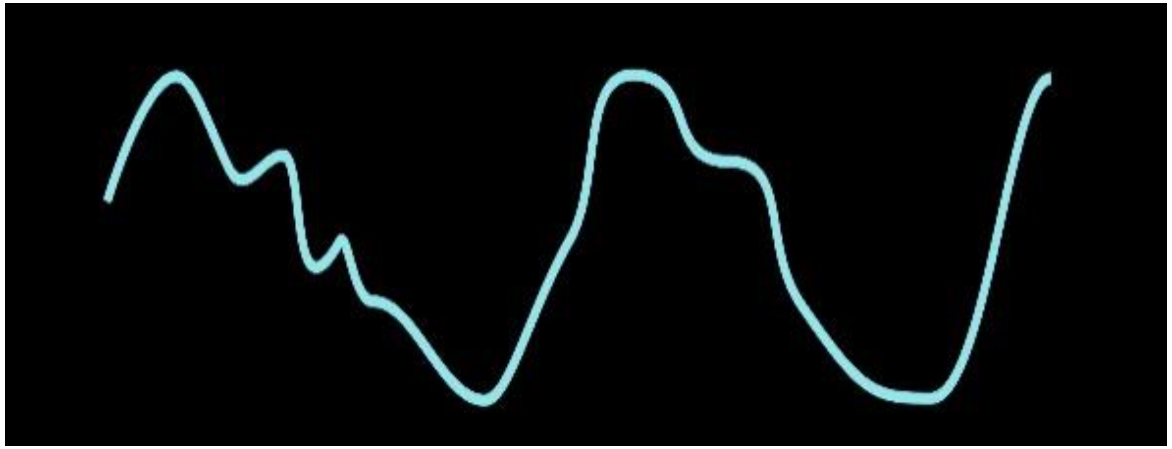
Note: for the time being we will focus on deterministic processes.

Q. Can we write processes as sums of periodic functions (frequencies)?

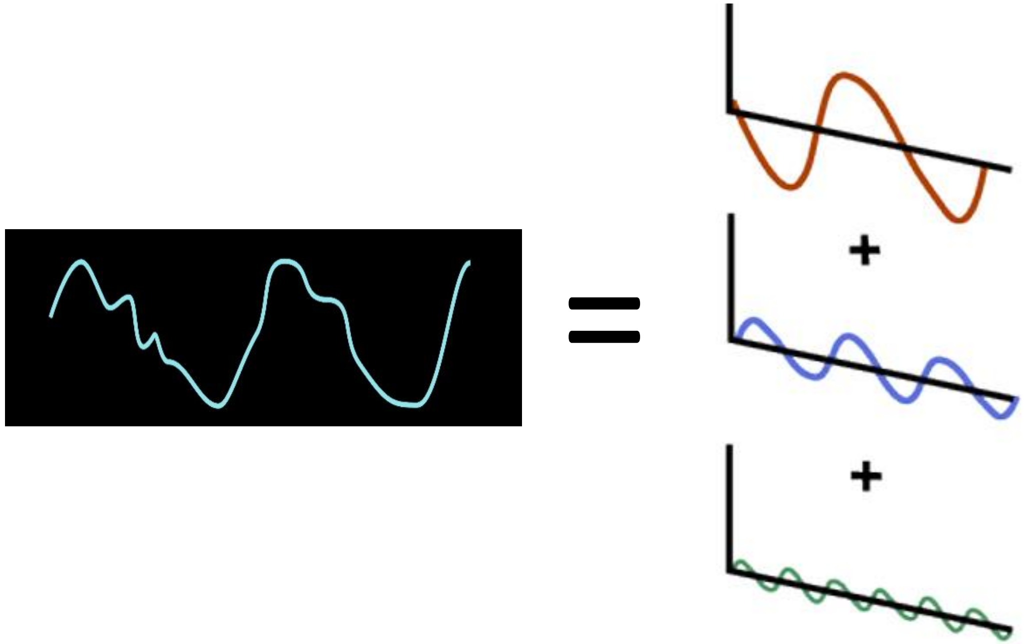


Baking a Fourier Cake

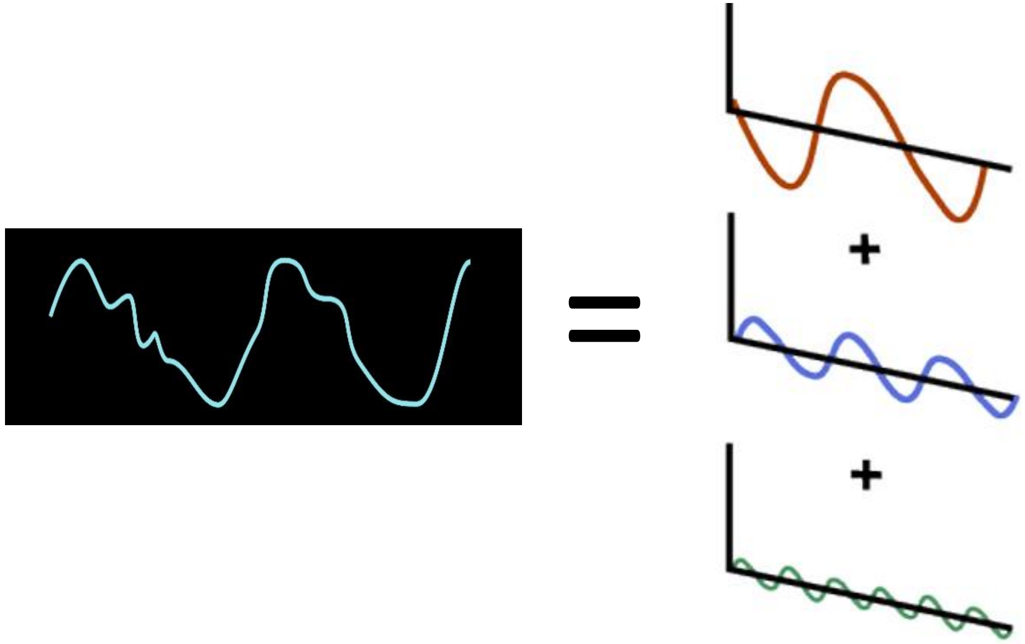
- **Given**: Signal shape (time-domain)
- **Ingredients**: Sinusoids of different frequencies
- **Choose**: How much of the each ingredient (sinusoid) to use?



- In Fourier Transform, we want to look at signals in terms of a fixed set of ingredients
    - Ingredients : Sinusoids of different frequencies

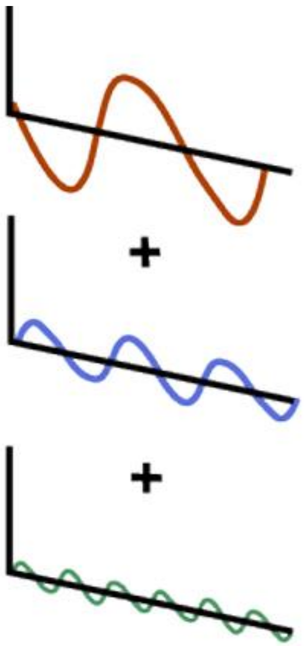


- In Fourier Transform, we want to look at signals in terms of a fixed set of ingredients
    - Ingredients : Sinusoids of different frequencies

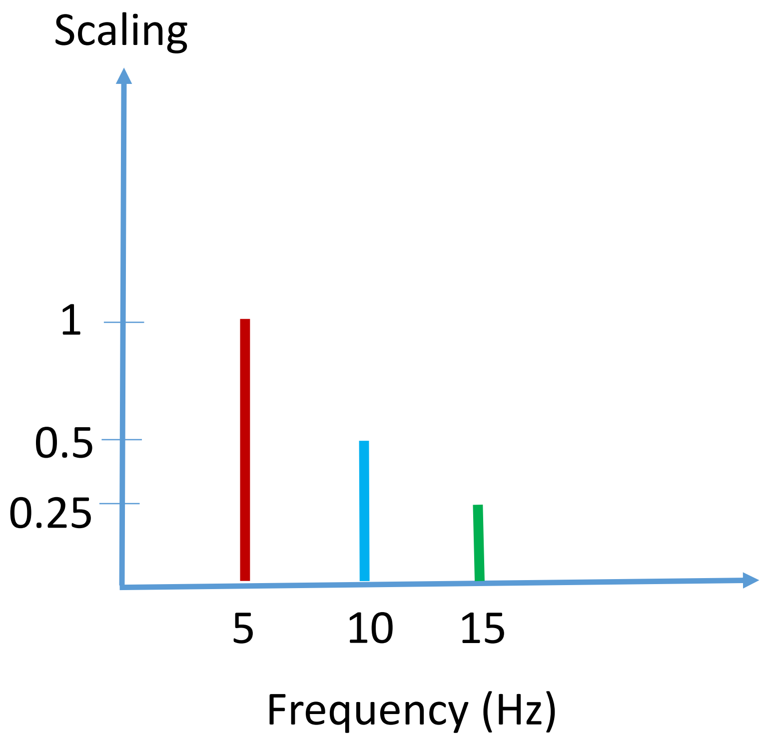


Ingredient (sinusoid frequency)	Amount (scaling)	Process
f_1	1	**Add all**
f_2	0.5	
f_3	0.25	

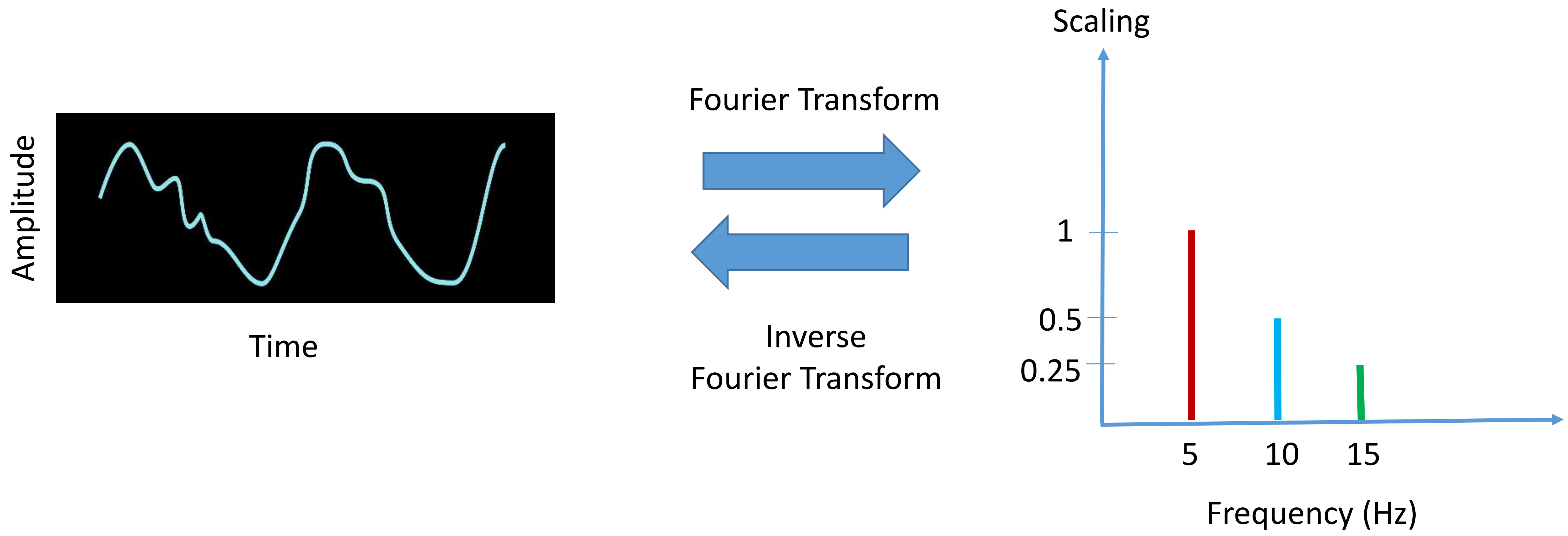
- How is this shown after Fourier transform?

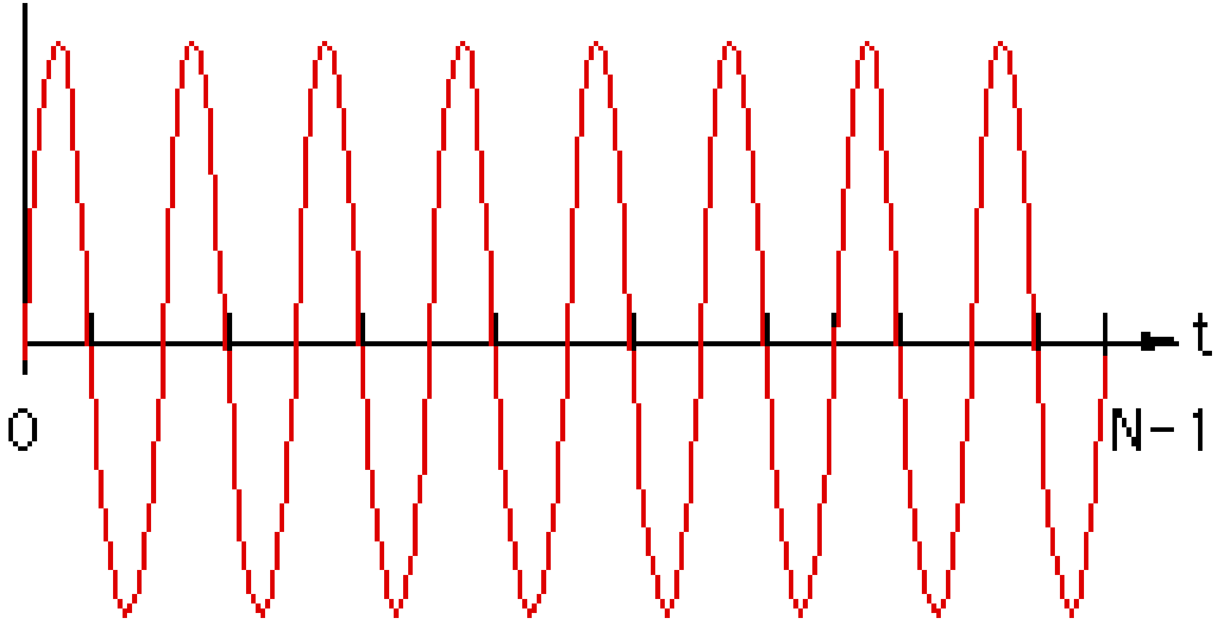


Ingredient (sinusoid frequency)	Amount (scaling)	Process
5 Hz	1	**Add all**
10 Hz	0.5	
15 Hz	0.25	



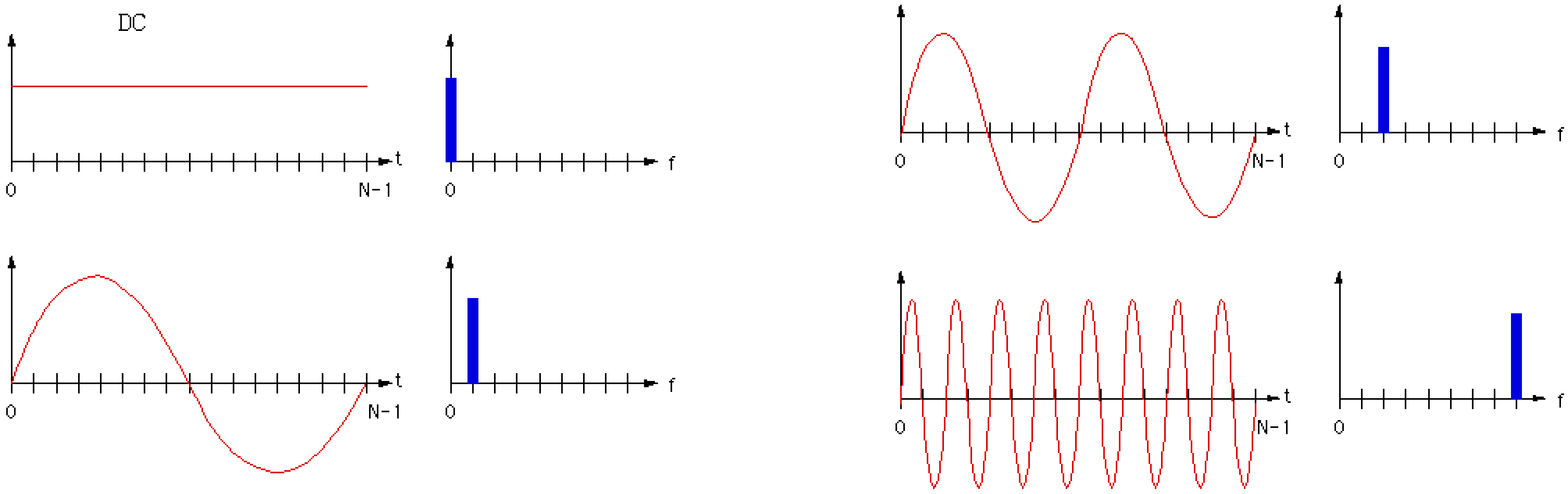
- We mostly skip the middle steps



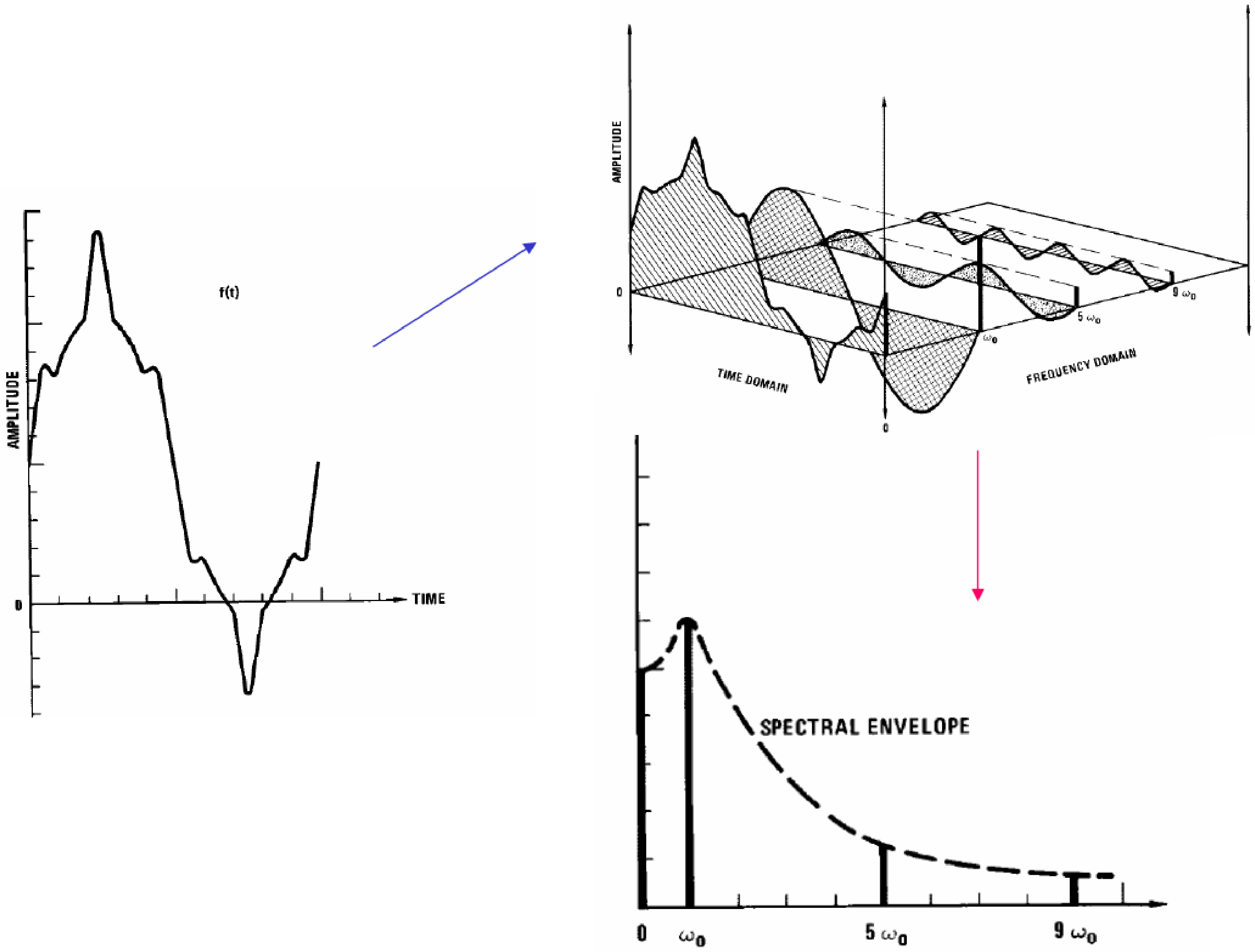


Fourier Transform = ?

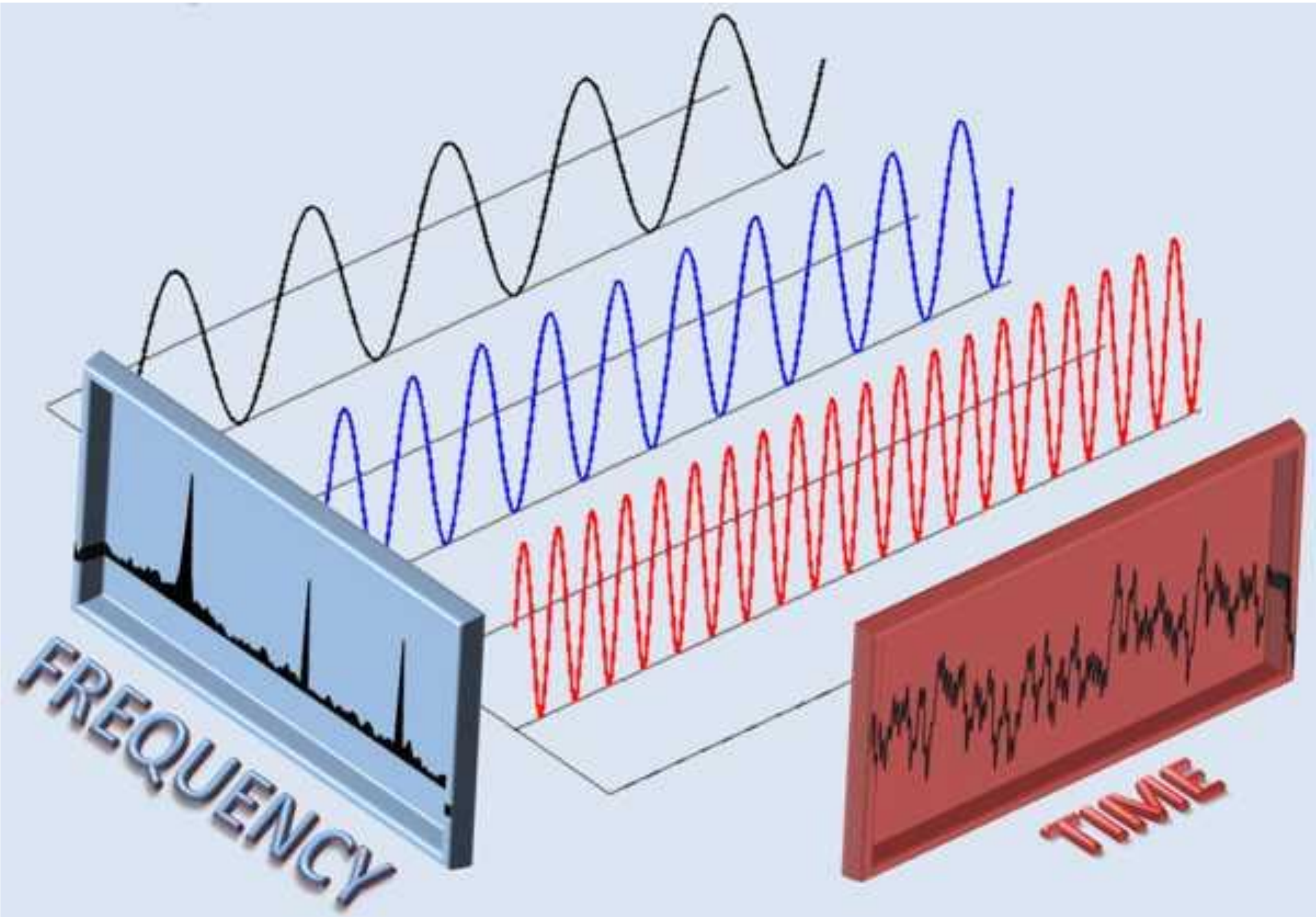
Some Fourier Transforms (Visual)

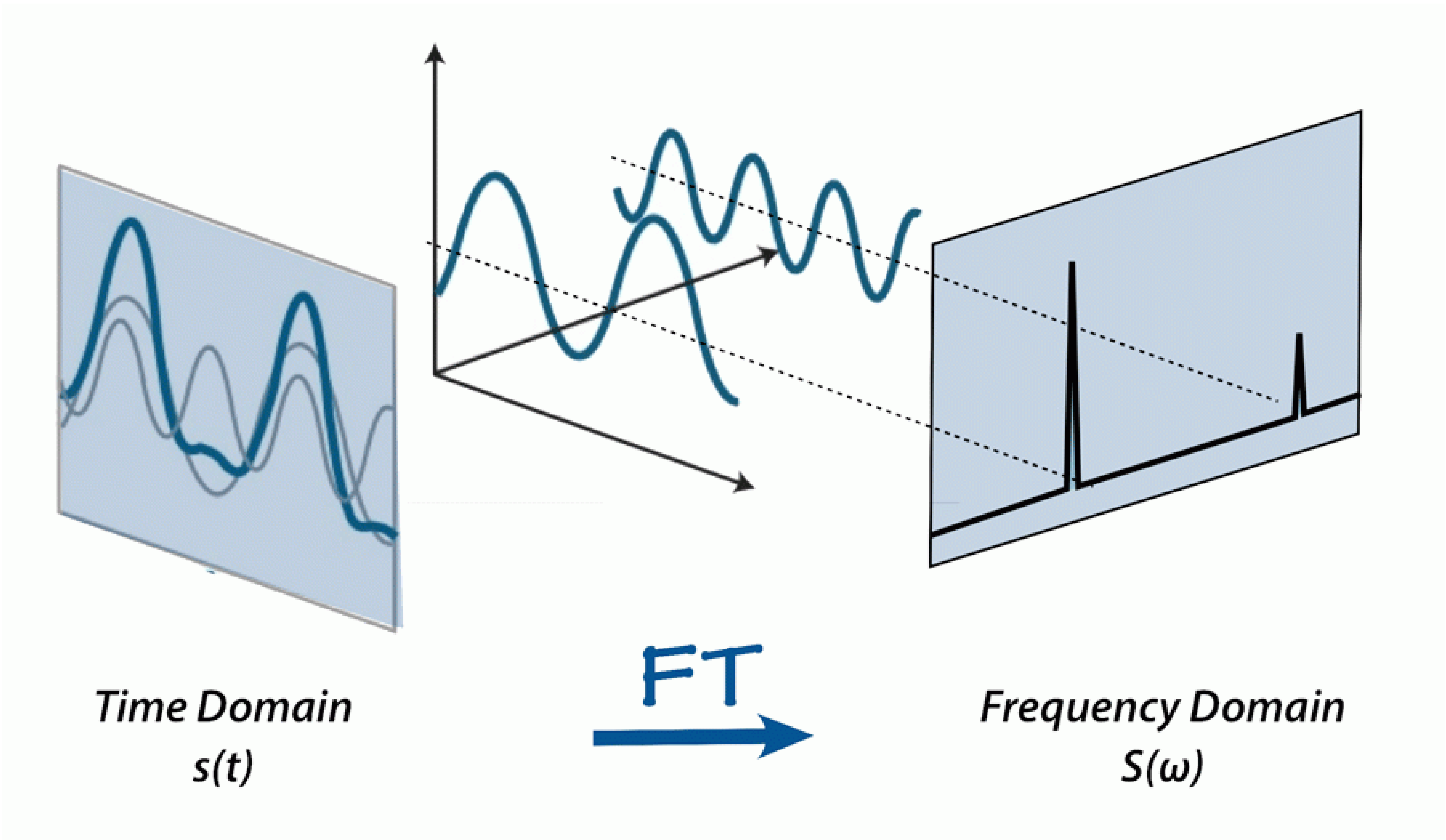


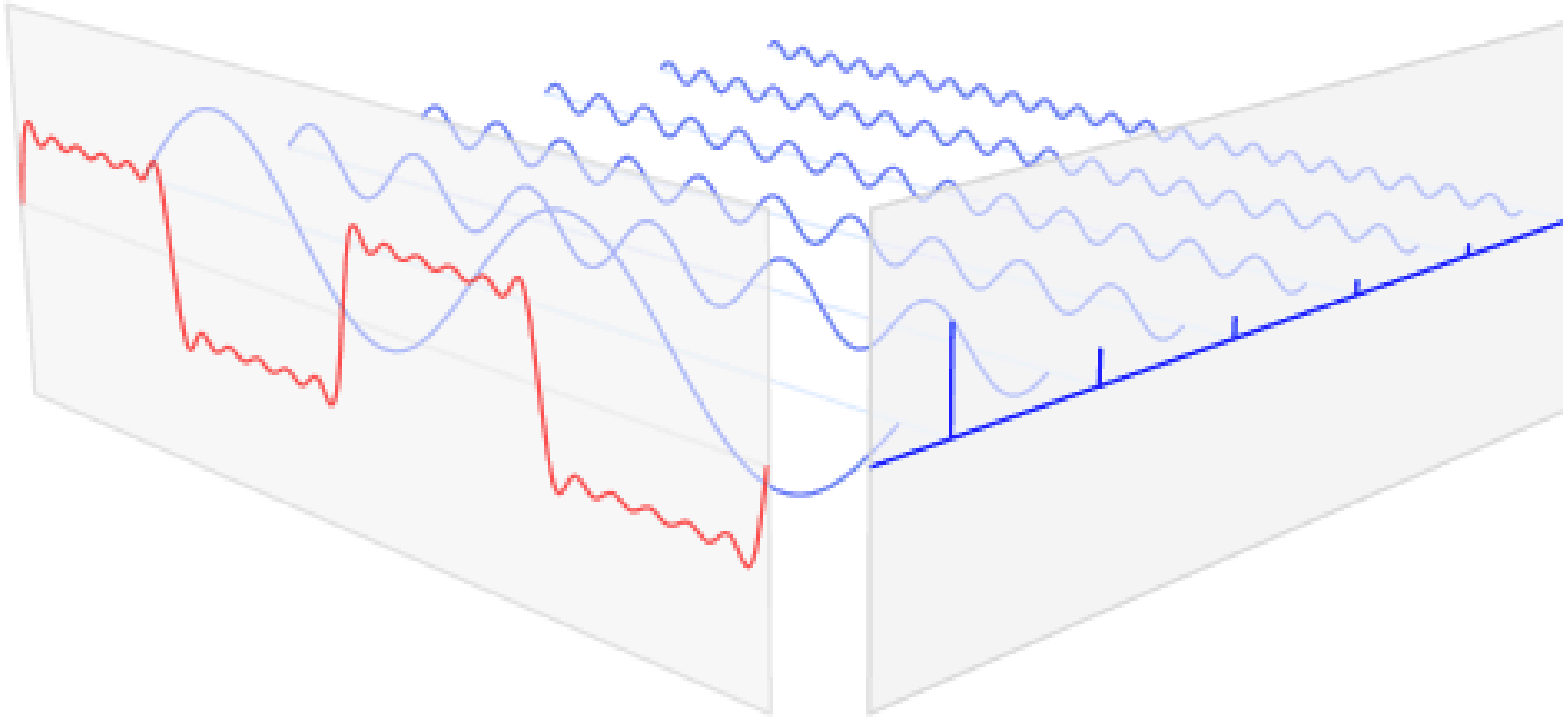
As frequencies increase, the FT peaks move outwards

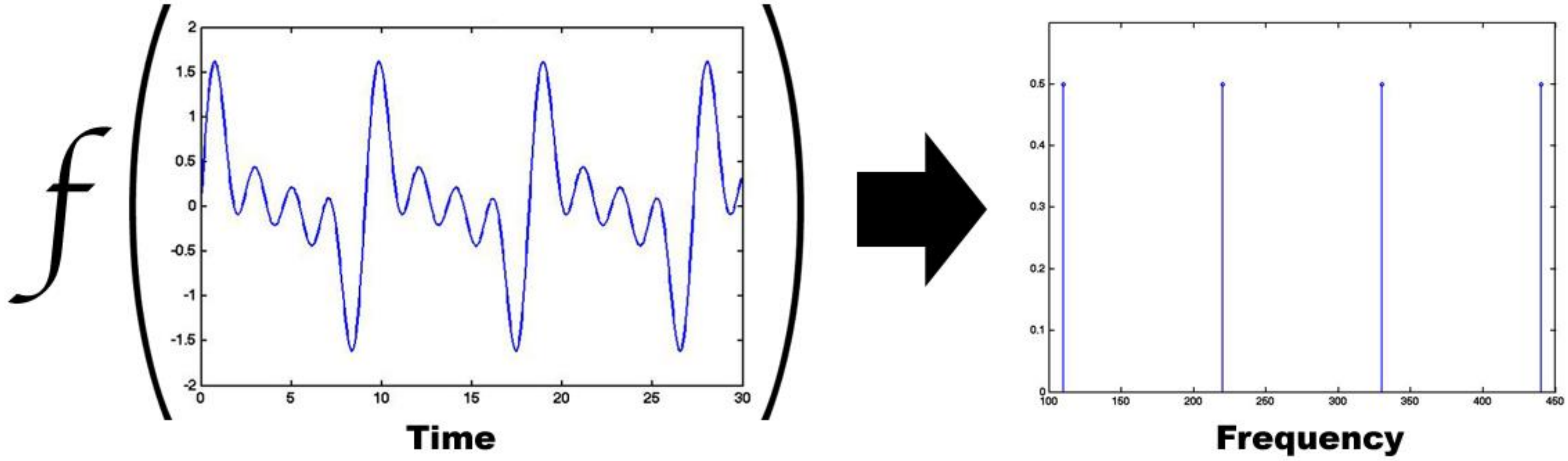


f(t)
AMPLITUDE
0
TIME
AMPLITUDE
0
TIME DOMAIN
FREQUENCY DOMAIN
ω_0
$5\ \omega_0$
$9\ \omega_0$
0
SPECTRAL ENVELOPE
0
ω_0
$5\ \omega_0$
$9\ \omega_0$

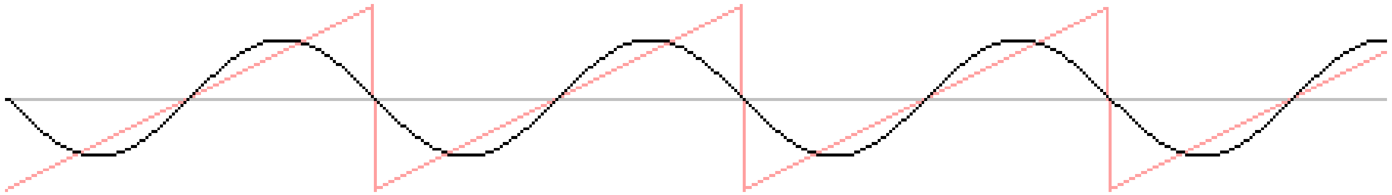

FREQUENCY
TIME

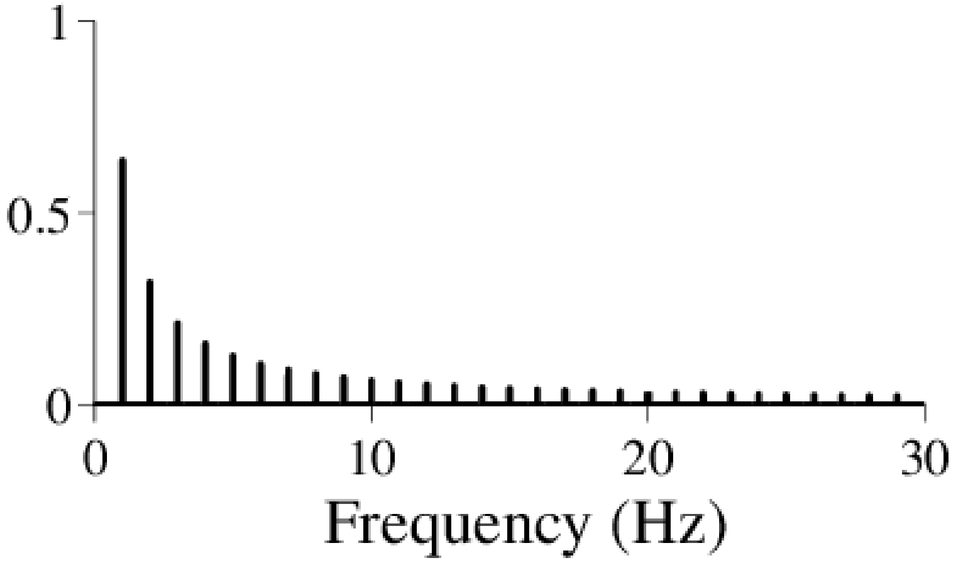

FT
Time Domain
s(t)
Frequency Domain
S(ω)

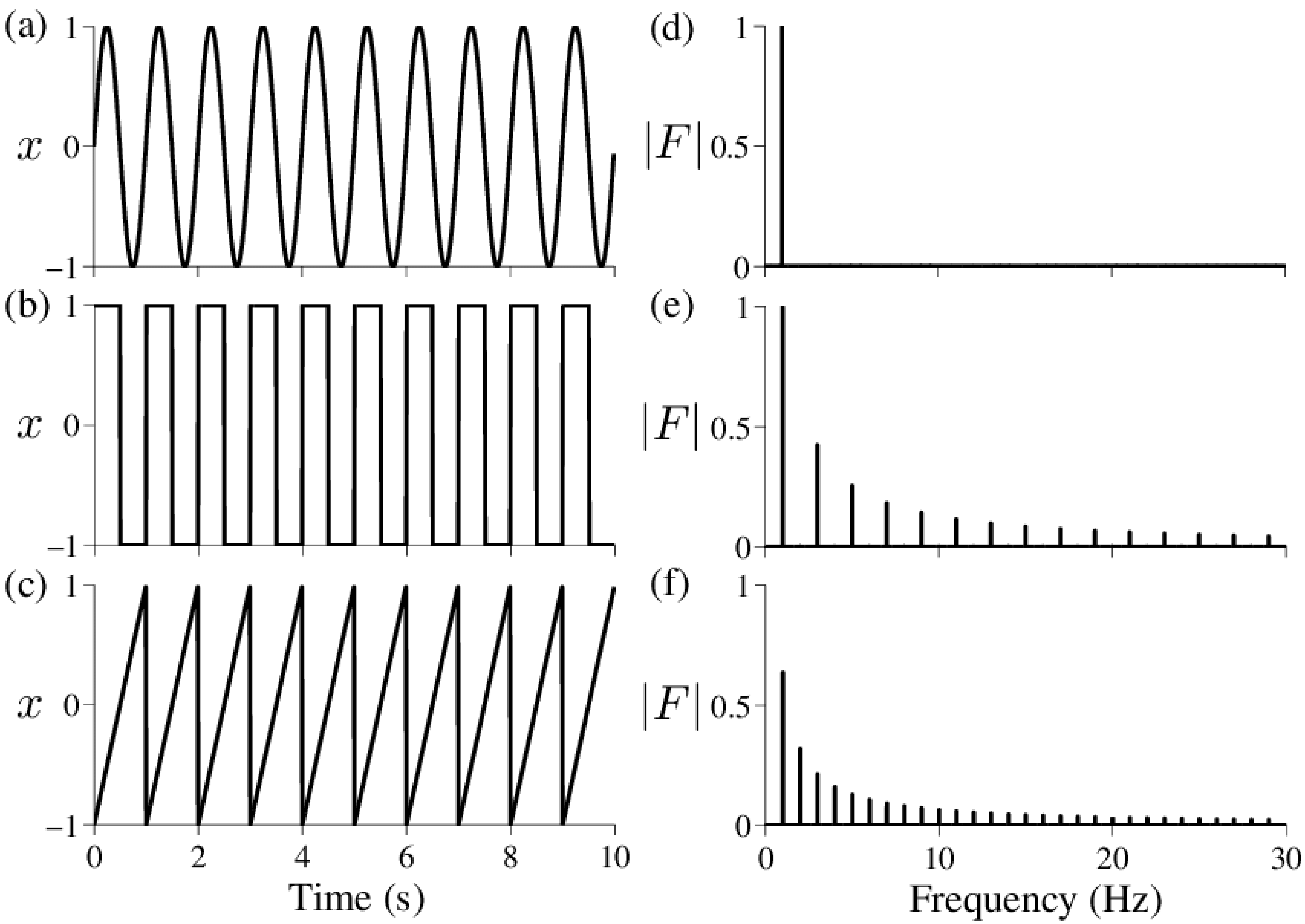



f
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
0
5
10
15
20
25
30
Time
0.5
0.4
0.3
0.2
0.1
0
100
150
200
250
300
350
400
450
Frequency

harmonics: 1






(a)
(b)
(c)
(d)
(e)
(f)
x
|F|
Time (s)
Frequency (Hz)

$f(x)$
1ucasvb.tumblr.com

Q. Is it always possible to write signals as sums of sinusoids?

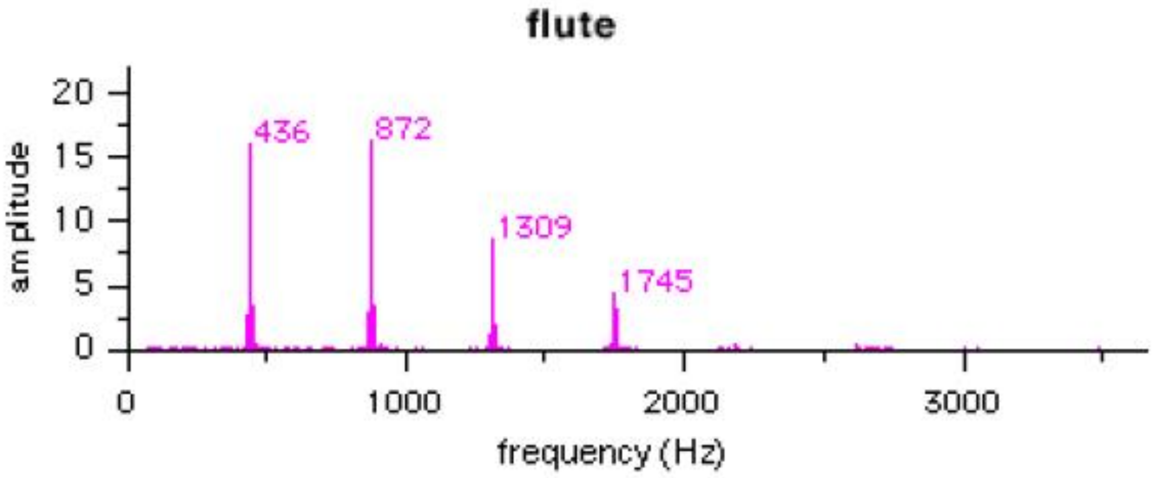
Q. Is it always possible to write signals as sums of sinusoids?

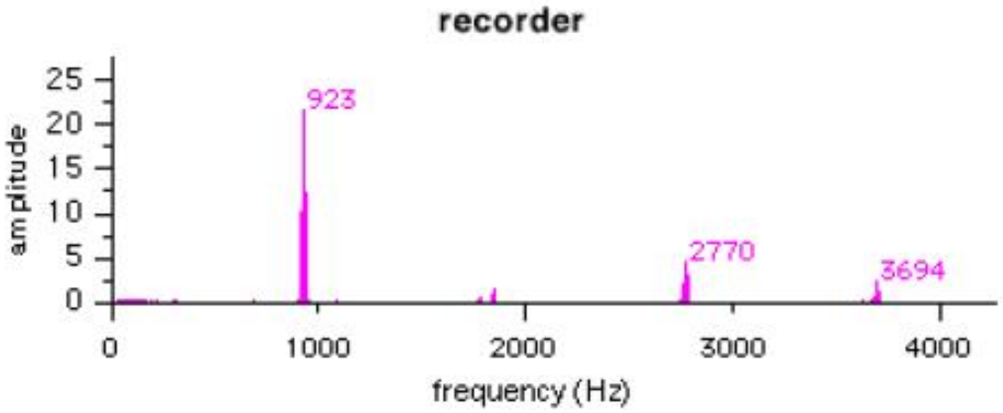
- **No**. There are theoretical signals that do not have a Fourier Transform (e.g., $e^{-at}u(t)$ with $a < 0$).

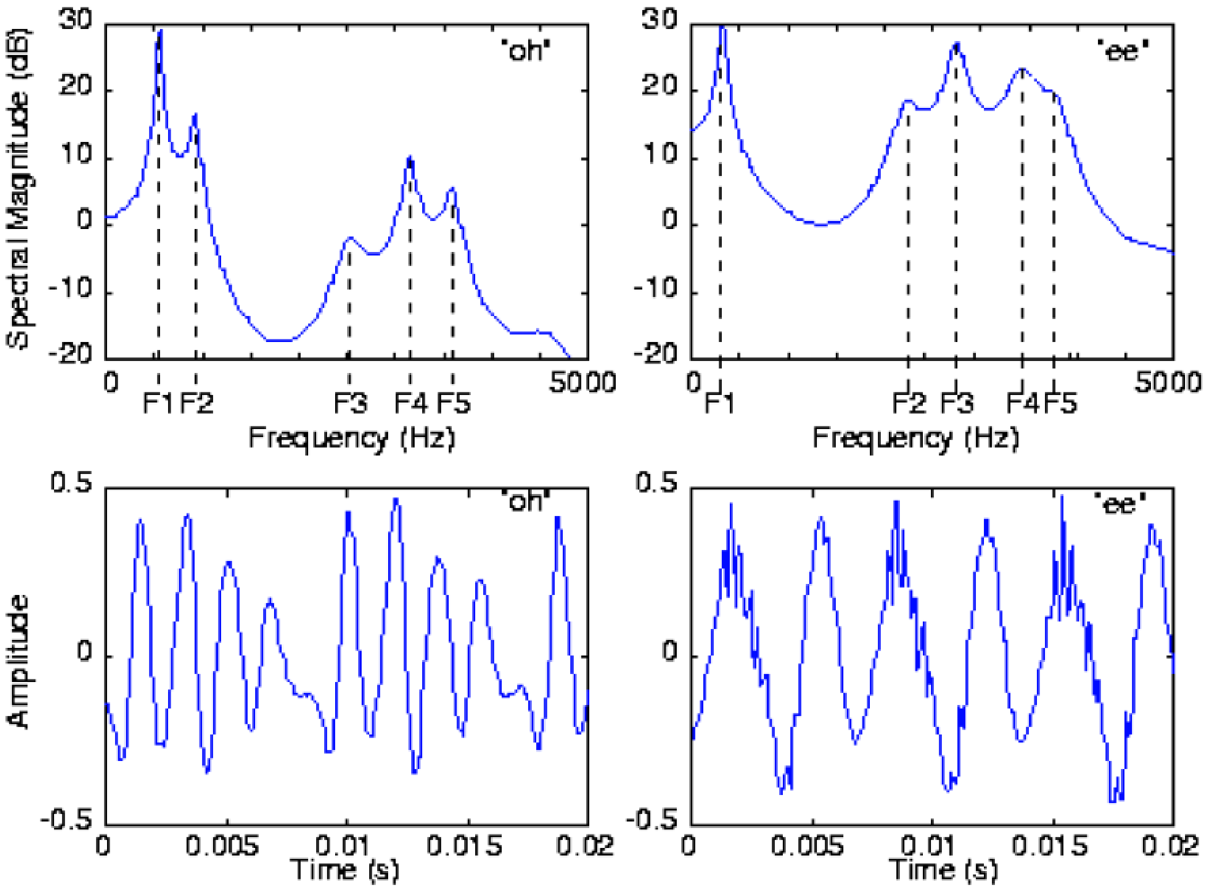
- **However**, all physically realizable signals have Fourier Transforms.

Q. Why Use Periodic Functions (frequencies)?

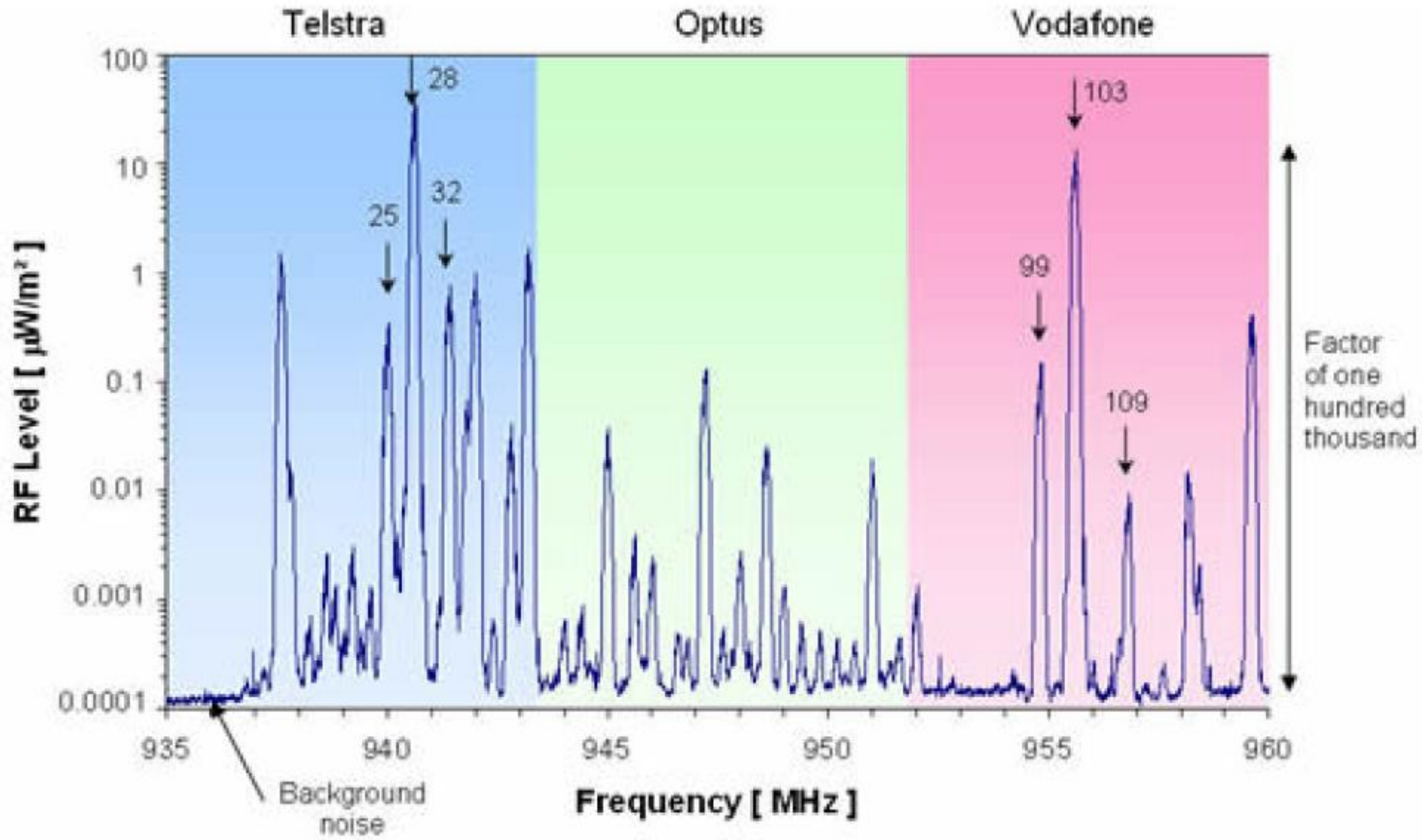
A large number of physical phenomena have underlying periodicities (frequencies)…


flute
amplitude
20
15
10
5
0
436
872
1309
1745
0
1000
2000
3000
frequency (Hz)

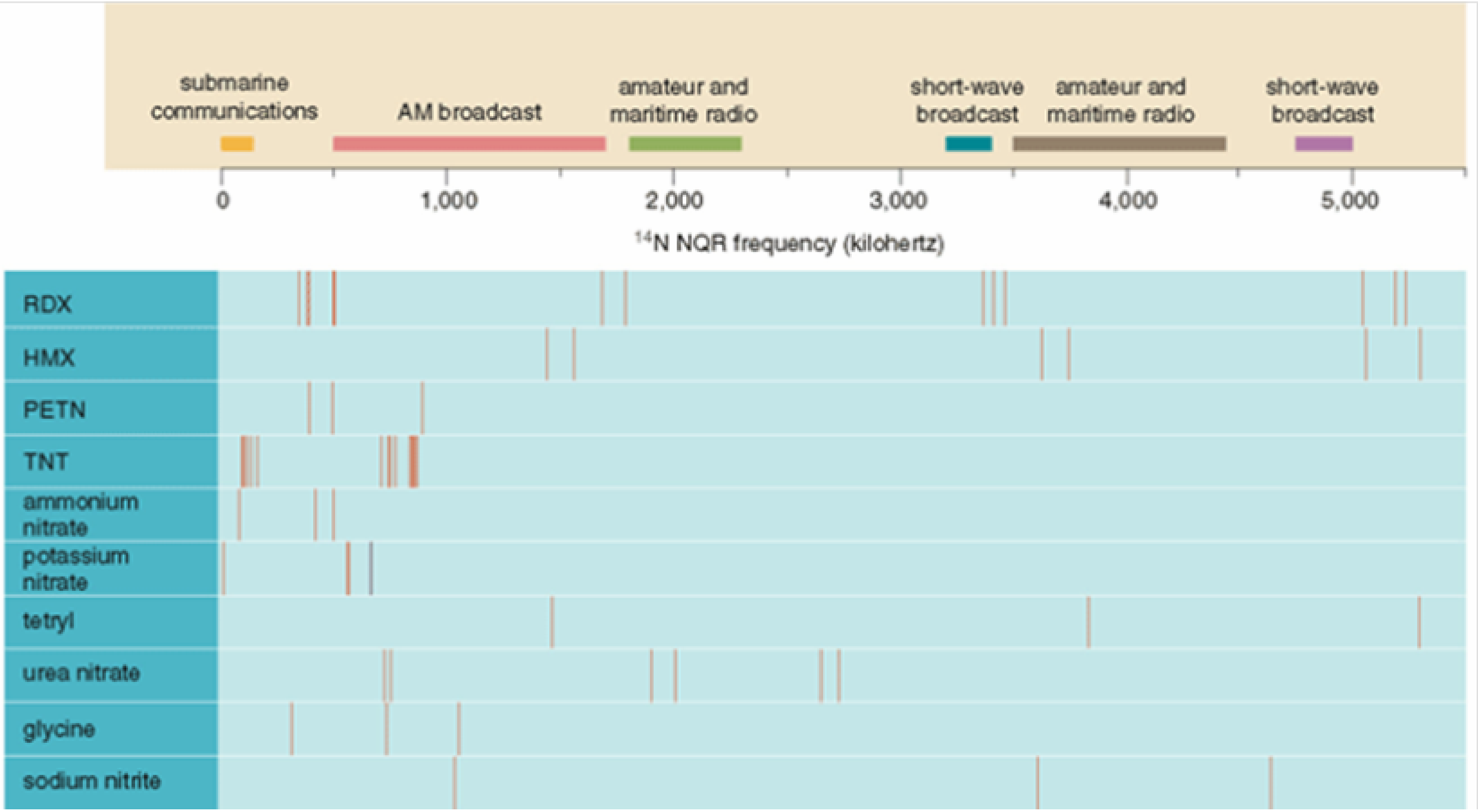

recorder
amplitude
25
20
15
10
5
0
923
2770
3694
0
1000
2000
3000
4000
frequency (Hz)



Voice Recognition

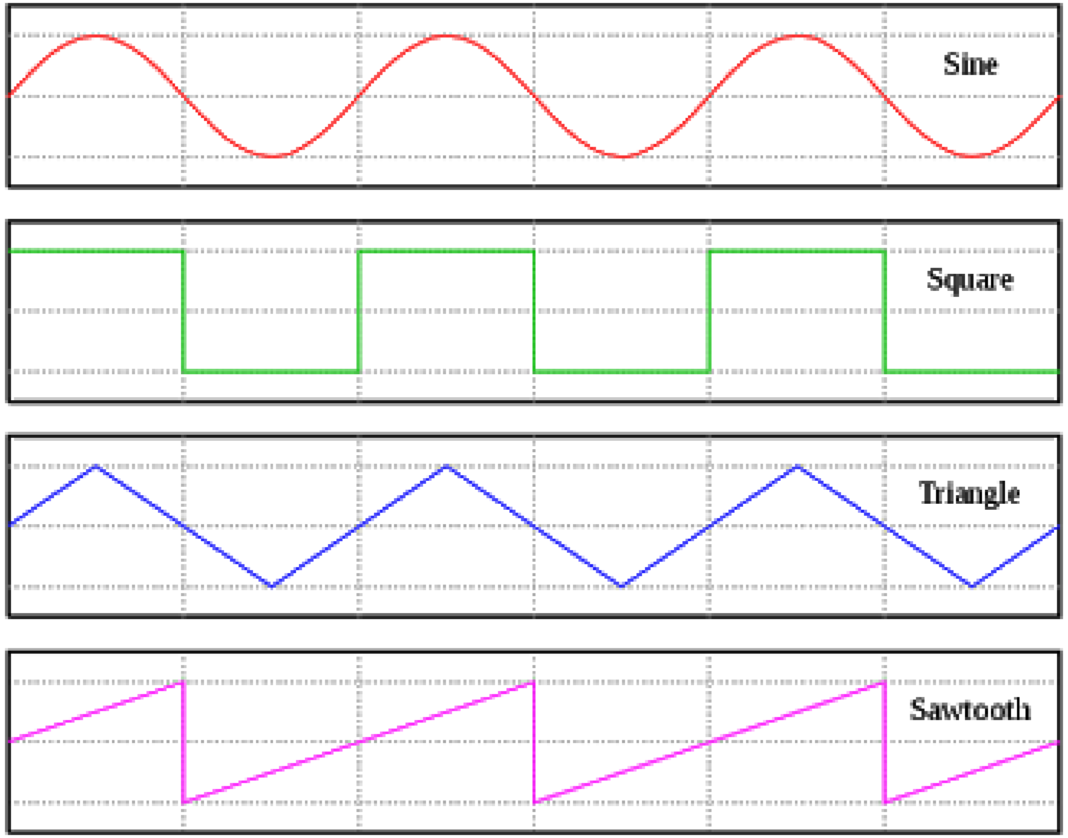


Communication systems are based on the frequencies of tunable antennas



Chemicals may be identified by the unique resonant frequencies of their nuclei or molecules

Q. Why Sinusoids?

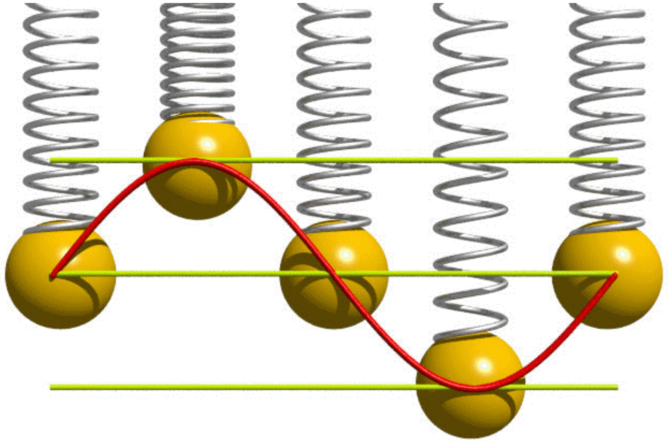


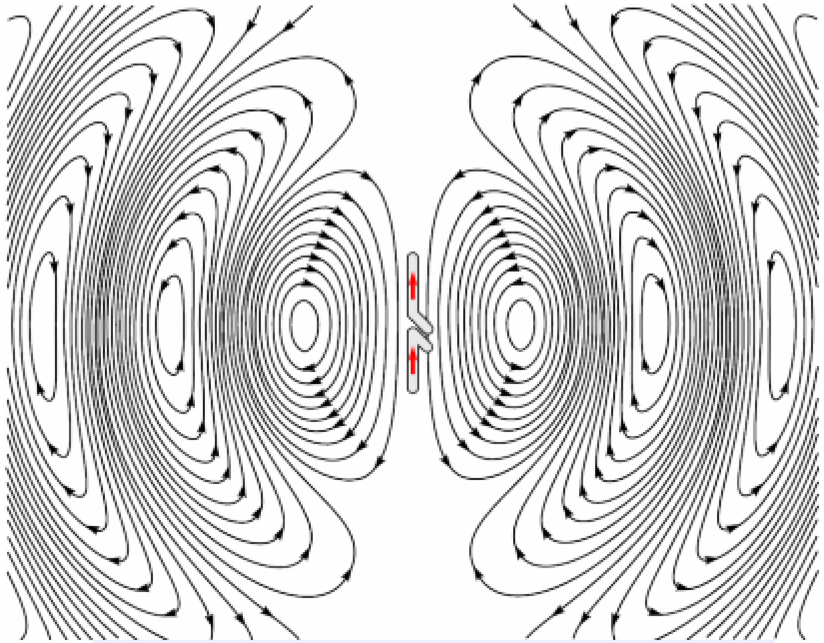
Why not other types of periodic functions?

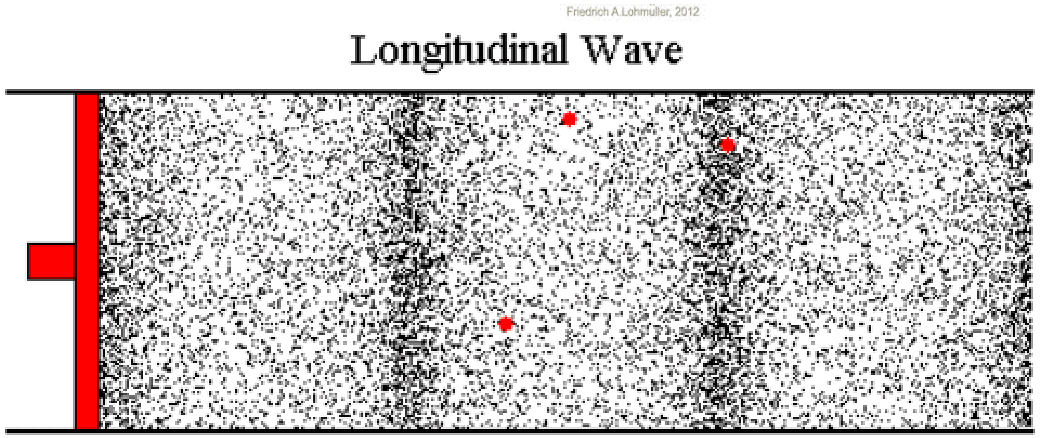
Q. Why Sinusoids?

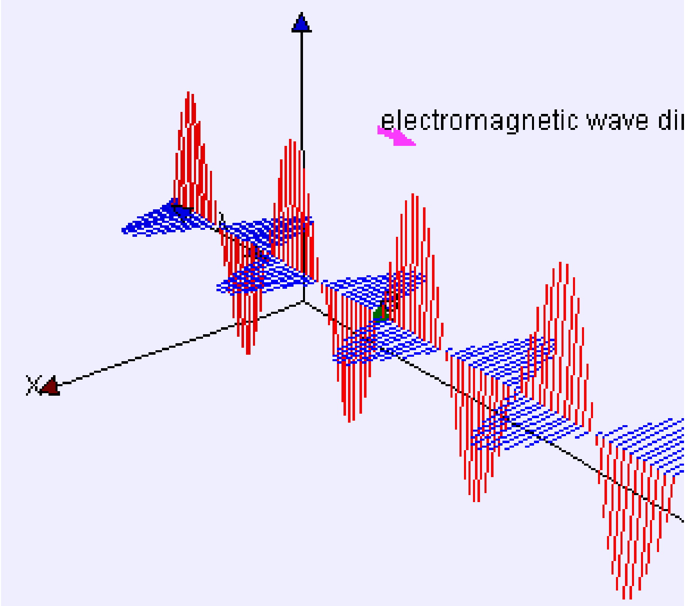
- **Smooth** (analytically simpler, e.g., differentiable, integrable…)
- Nicely reflect behavior of **natural phenomena** (to-and-fro motions)
- Sinusoids are Eigen-functions of linear time-invariant operators (which helps simplify computations)

Q. Why Sinusoids?









In fact, Fourier Transform does not use just sinusoids, it uses ***complex sinusoids****!!!*

Why?

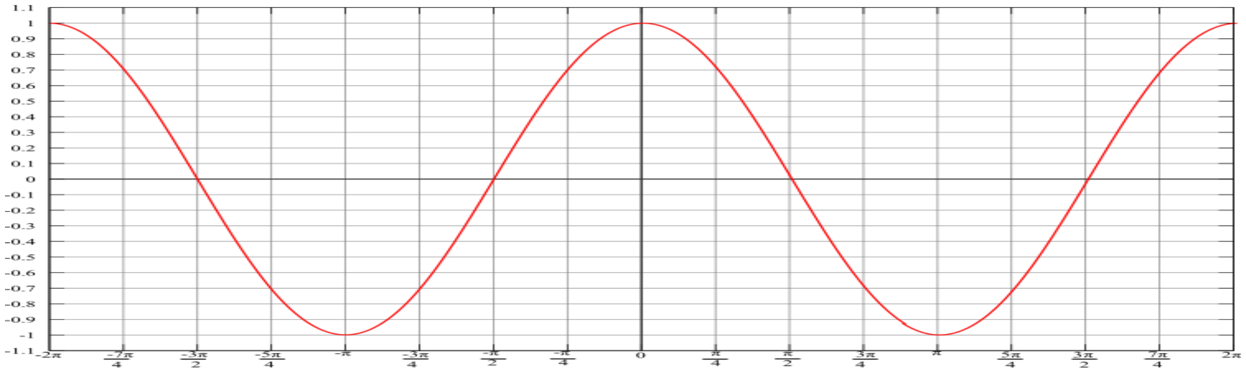
- More general than real sinusoids
- More elegant analytically and in calculations

And what exactly was a complex sinusoid?

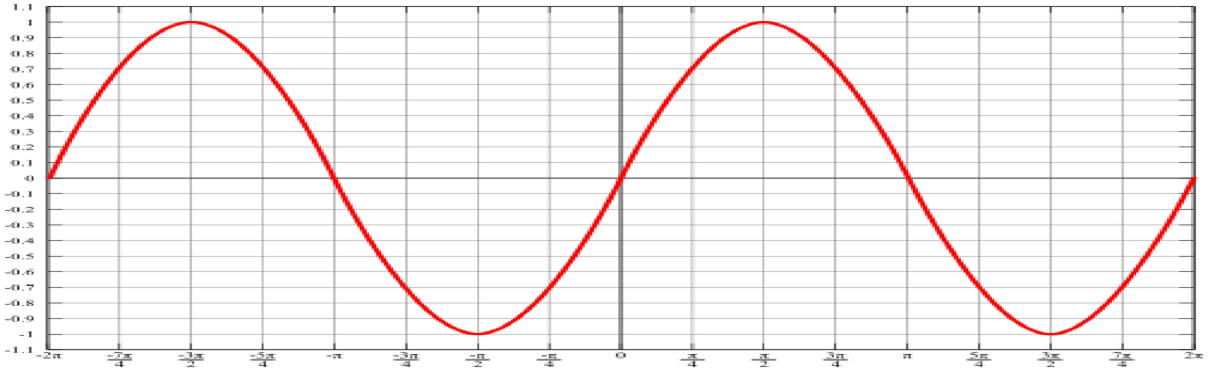
A Sine and a Cosine Walk Into an Imaginary Bar…

The Complex Sinusoid - $e^{j\omega t}$

$\cos(\omega t)$

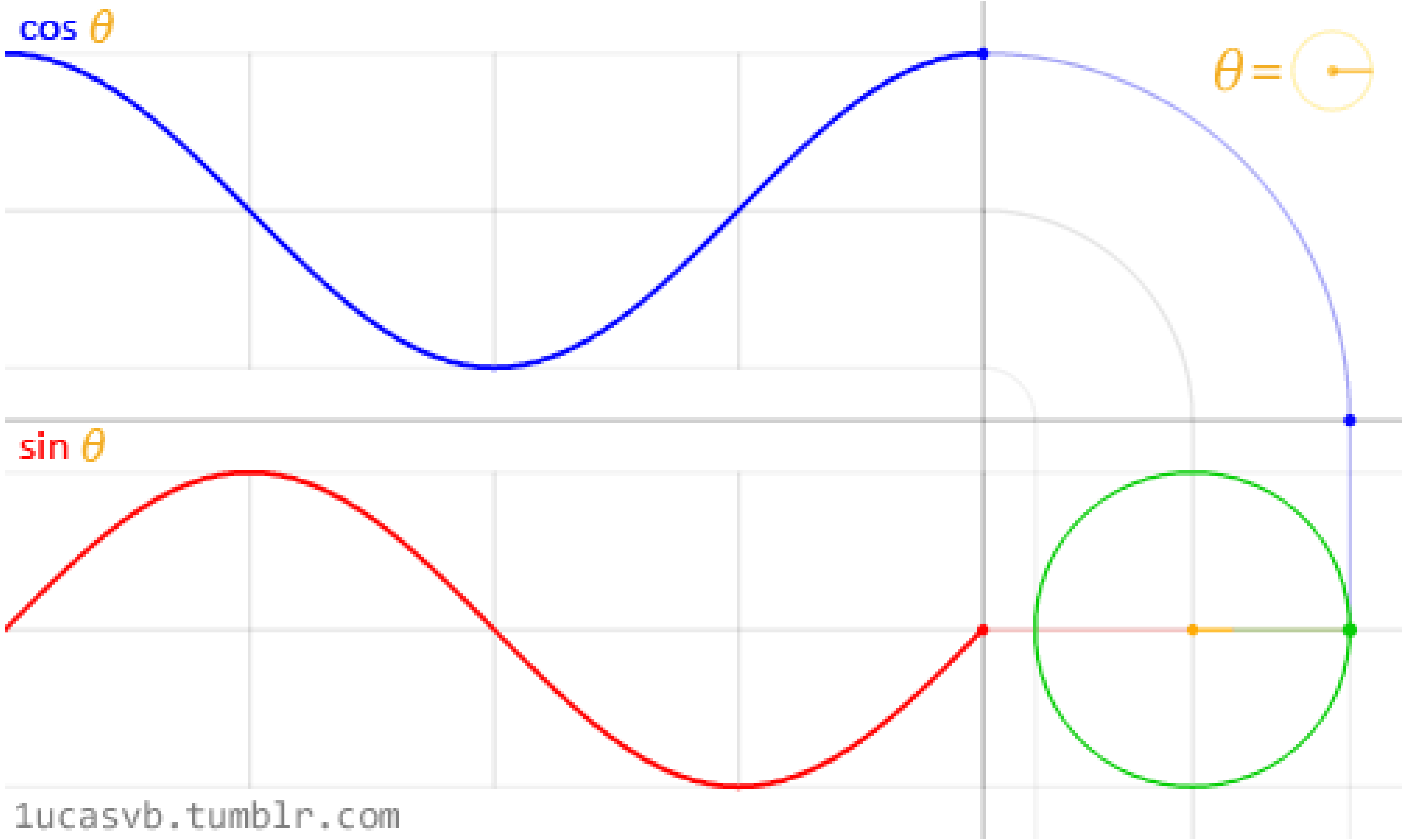


$\sin(\omega t)$

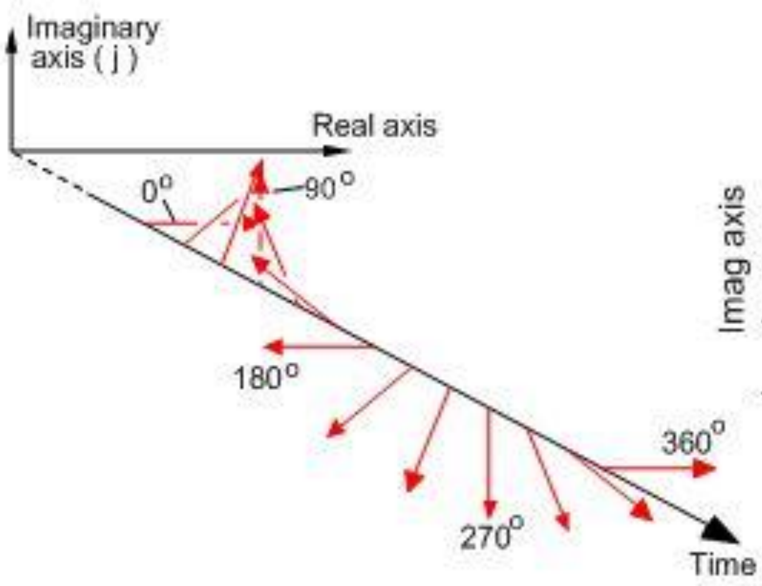


The Complex Sinusoid - $e^{j\omega t}$

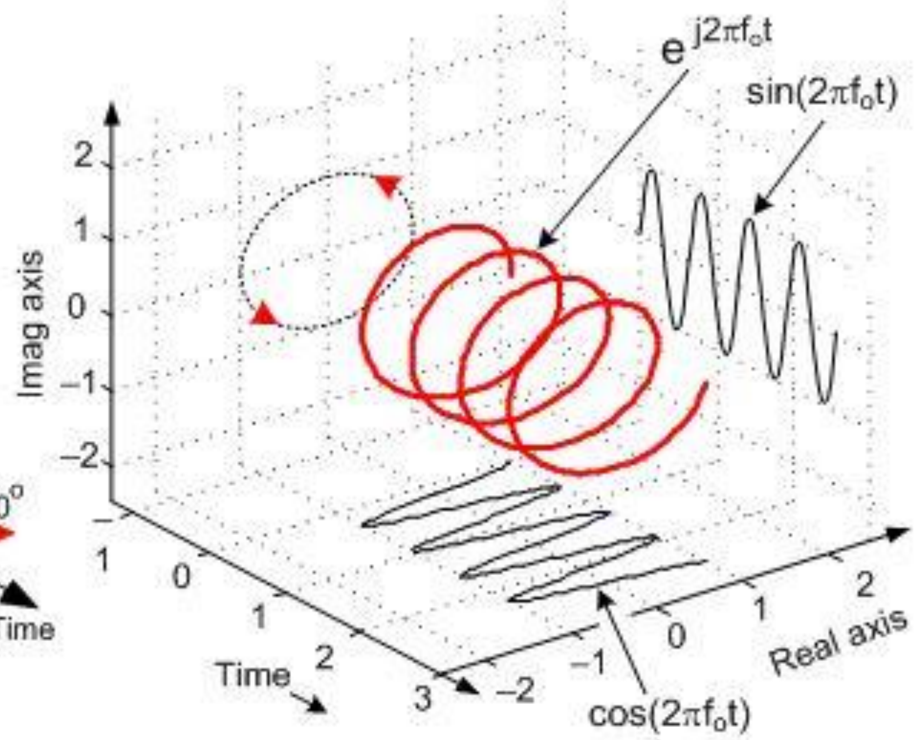
$$\cos(\omega t) + j\sin(\omega t)$$



The Complex Sinusoid - $e^{j\omega t} = \cos(\omega t) + j\sin(\omega t)$

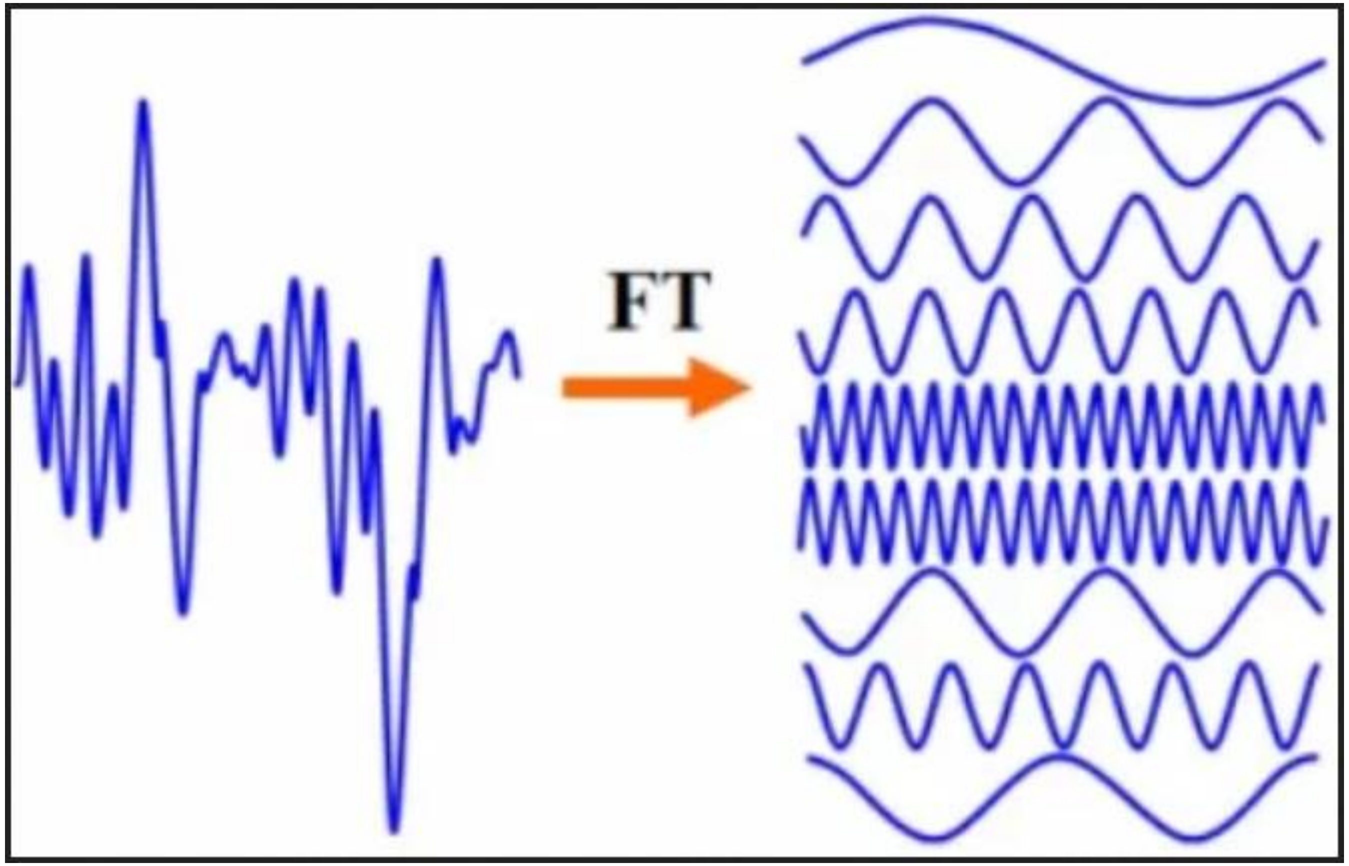


(a)



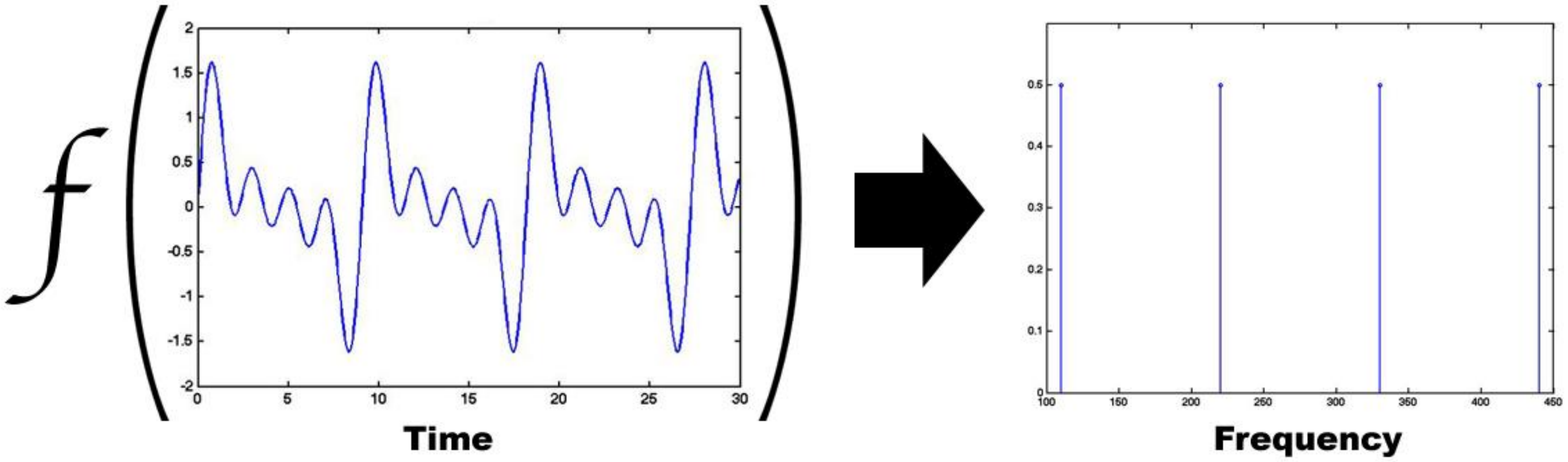
(b)

Mathematically Speaking…



$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega)e^{j\omega t}\,d\omega$$

Mathematically Speaking…

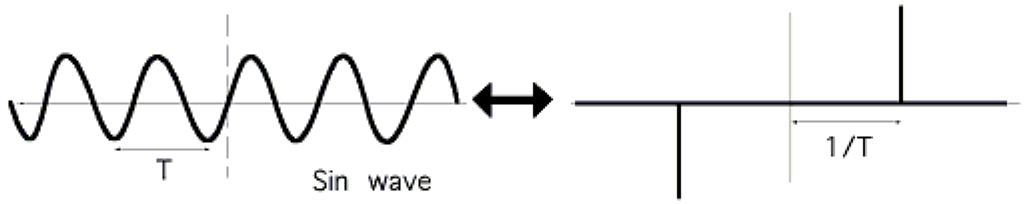


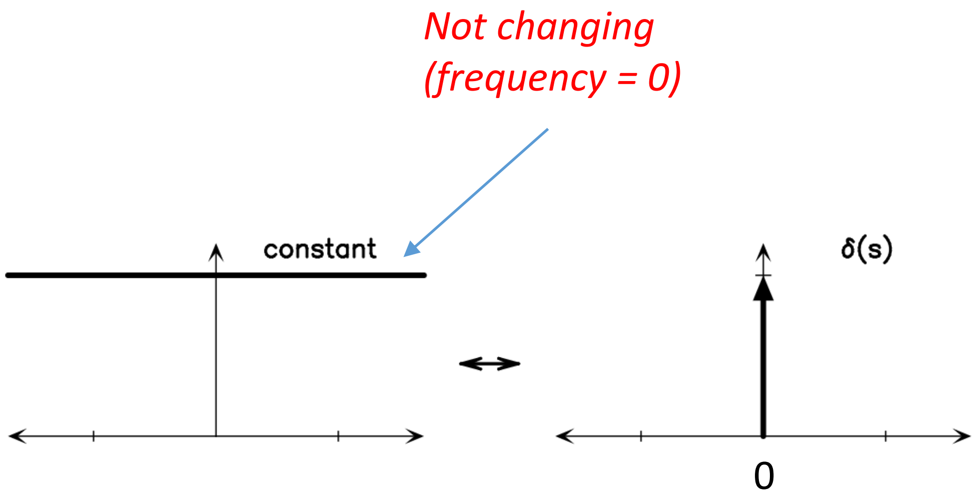
$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}dt$$

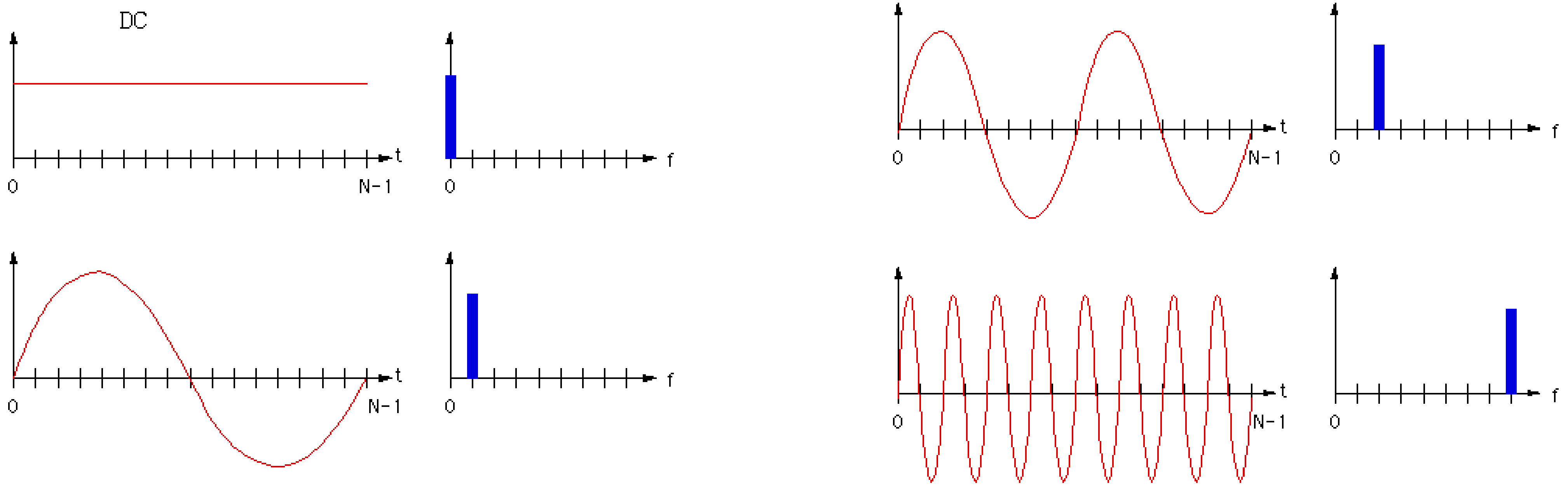
FT – Pairs and Rules-of-Thumb

Some Fourier Transforms (Visual)

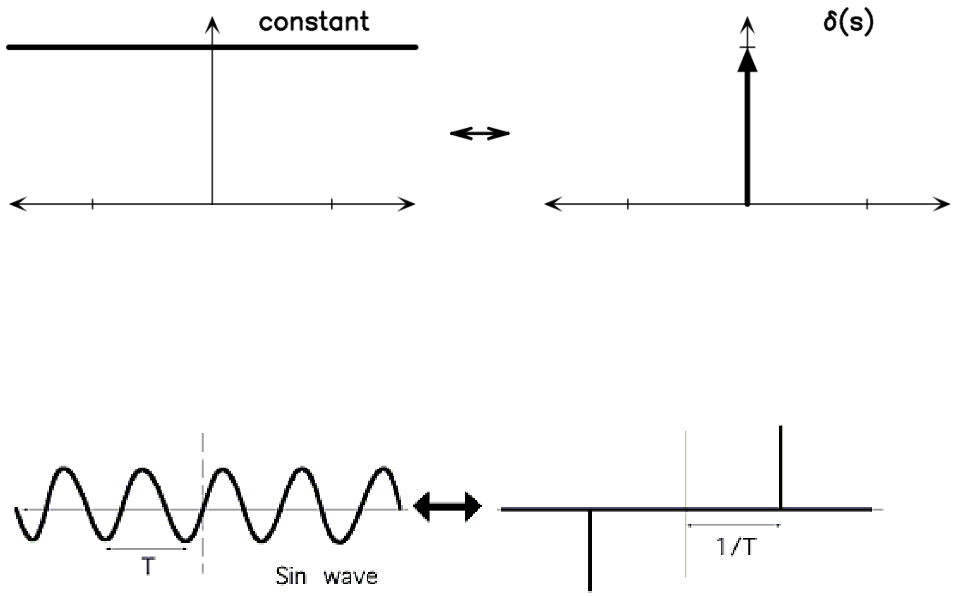
$$\sin(\omega t) = \frac{1}{2j}\left(e^{j\omega t} - e^{-j\omega t}\right)$$

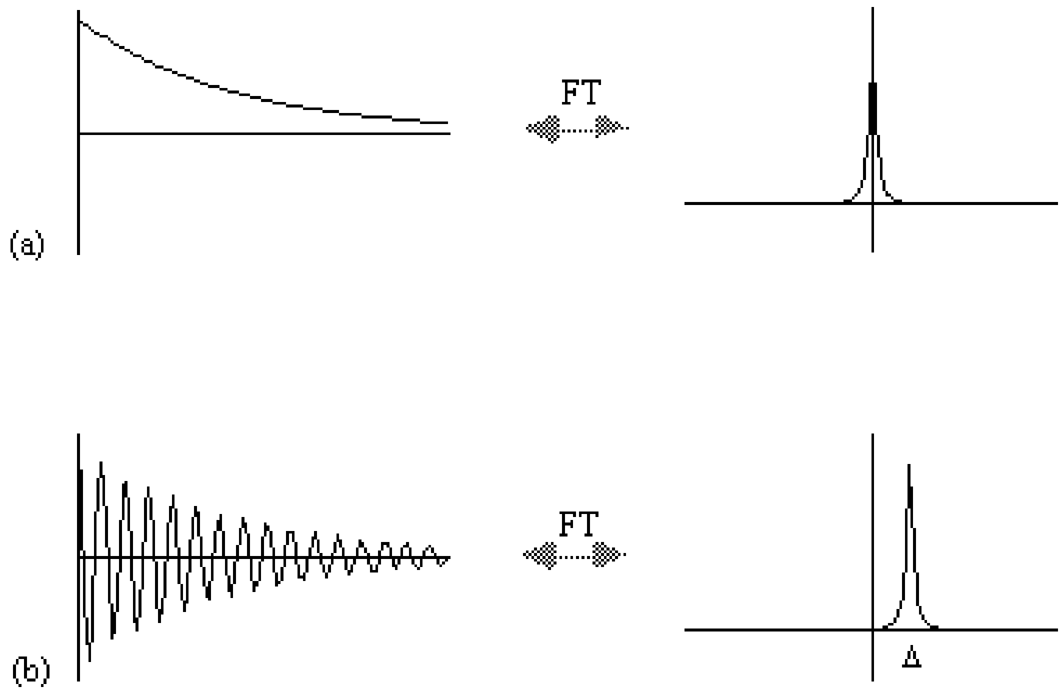




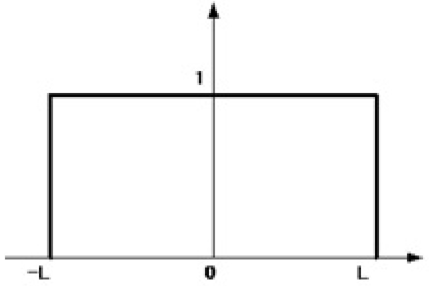


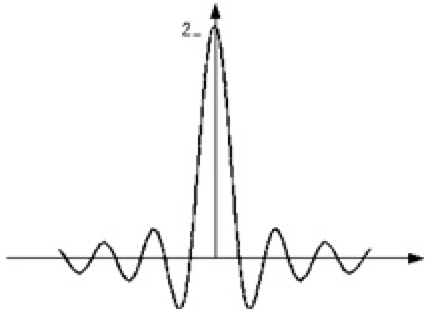
Rule1: As frequency increases, the FT peaks move outwards

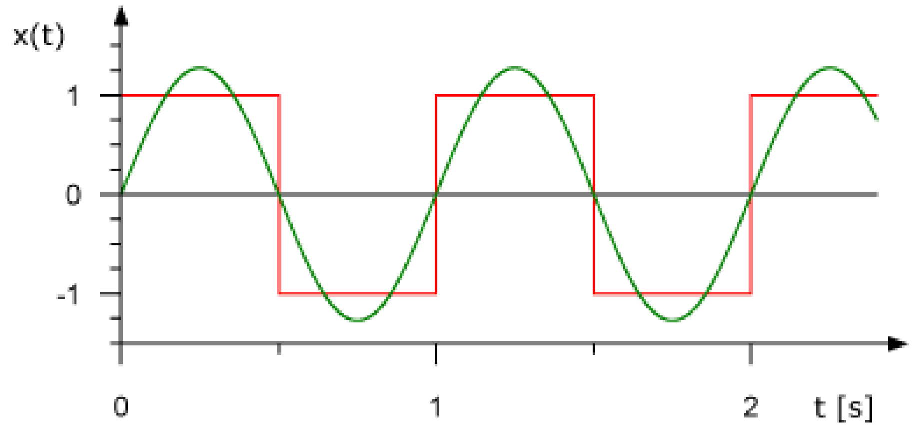


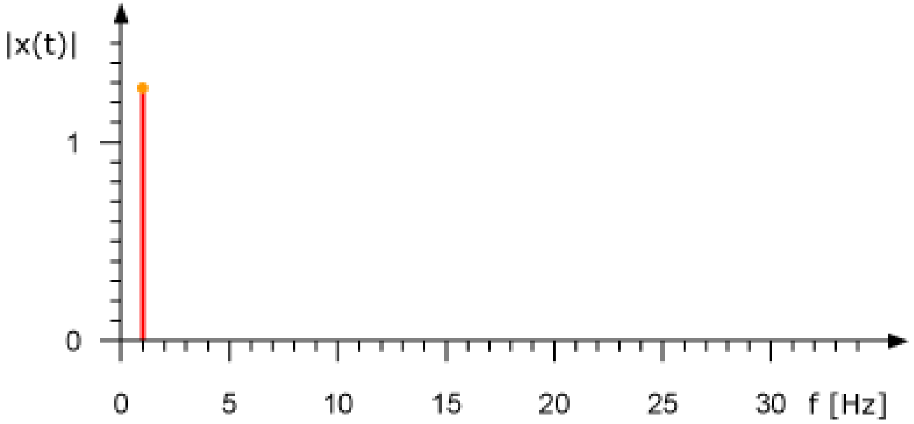


Rule 2: Damping causes spread

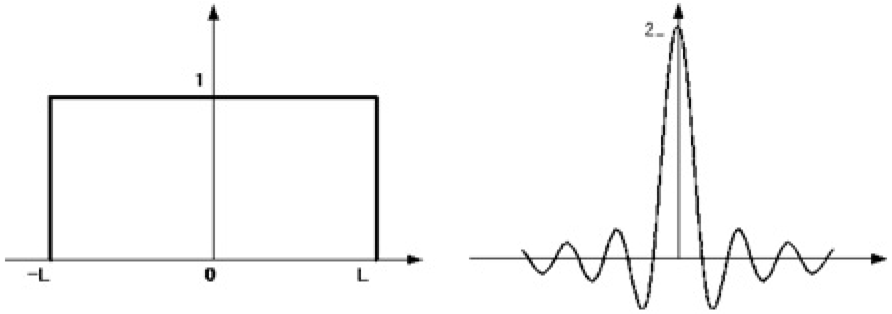


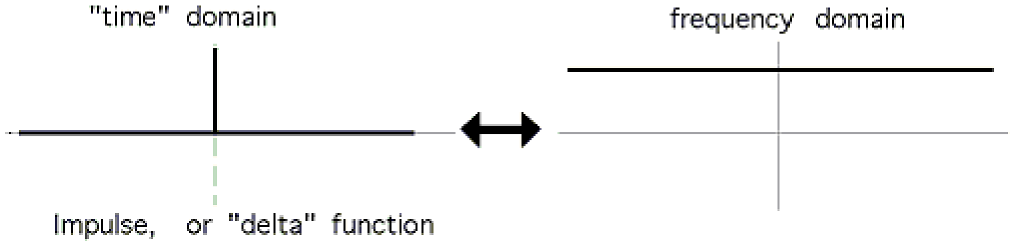






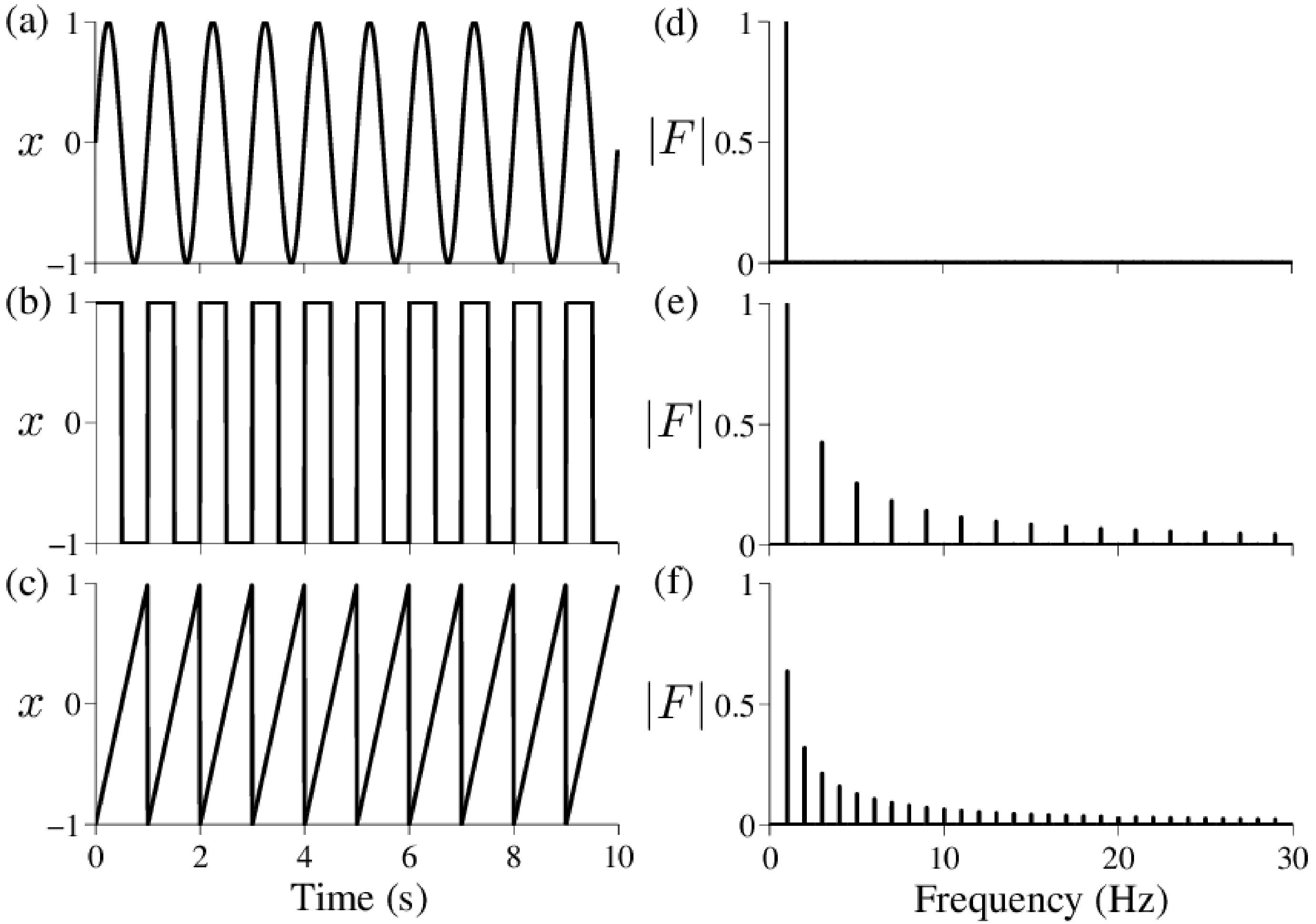
Rule 3: Sharp changes (edges) require a lot of frequencies



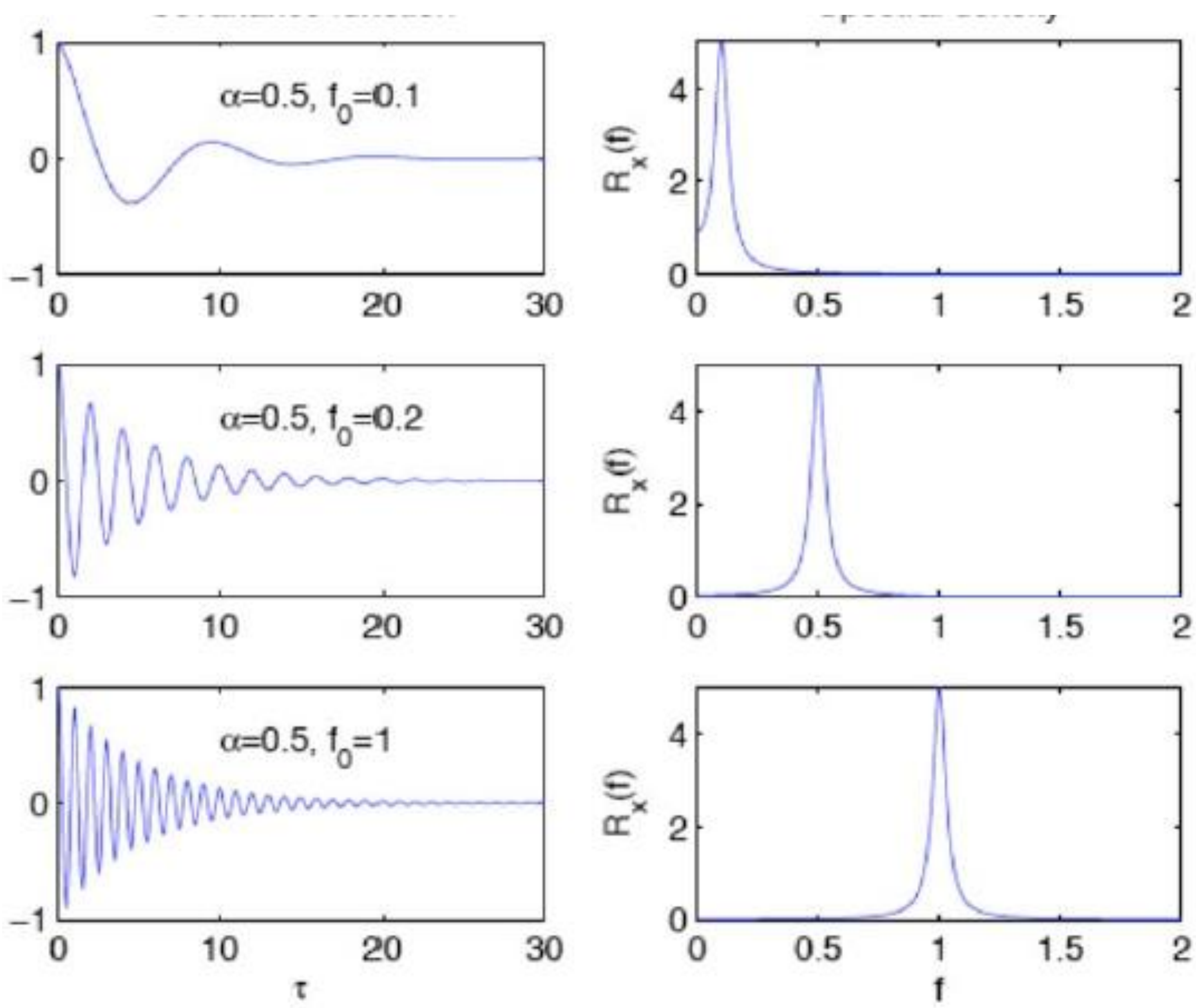


... and an extremely sharp change (impulse) requires ALL the frequencies!!

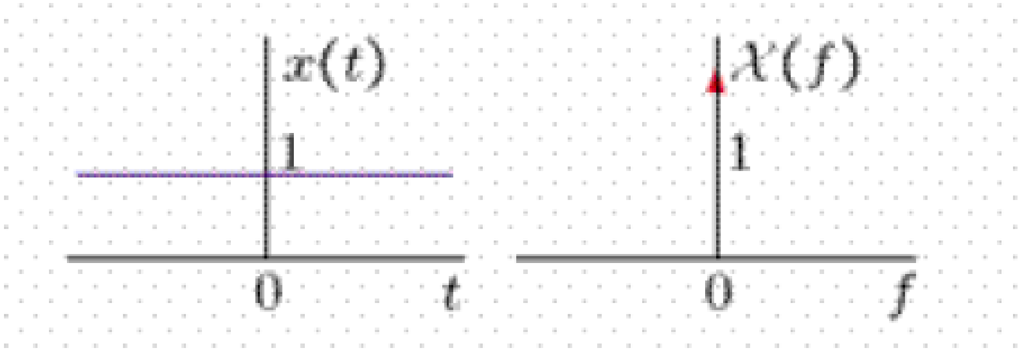
Rule 3: Sharp changes (edges) require a lot of frequencies

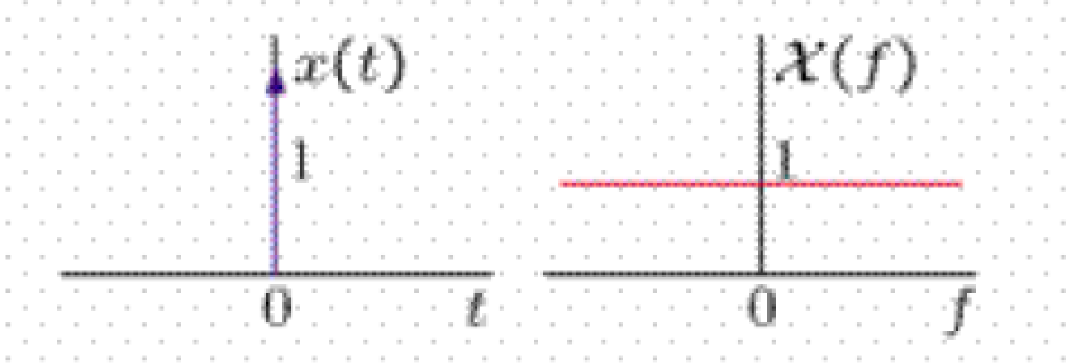


Rule 4: Periodic functions have discrete spectra.

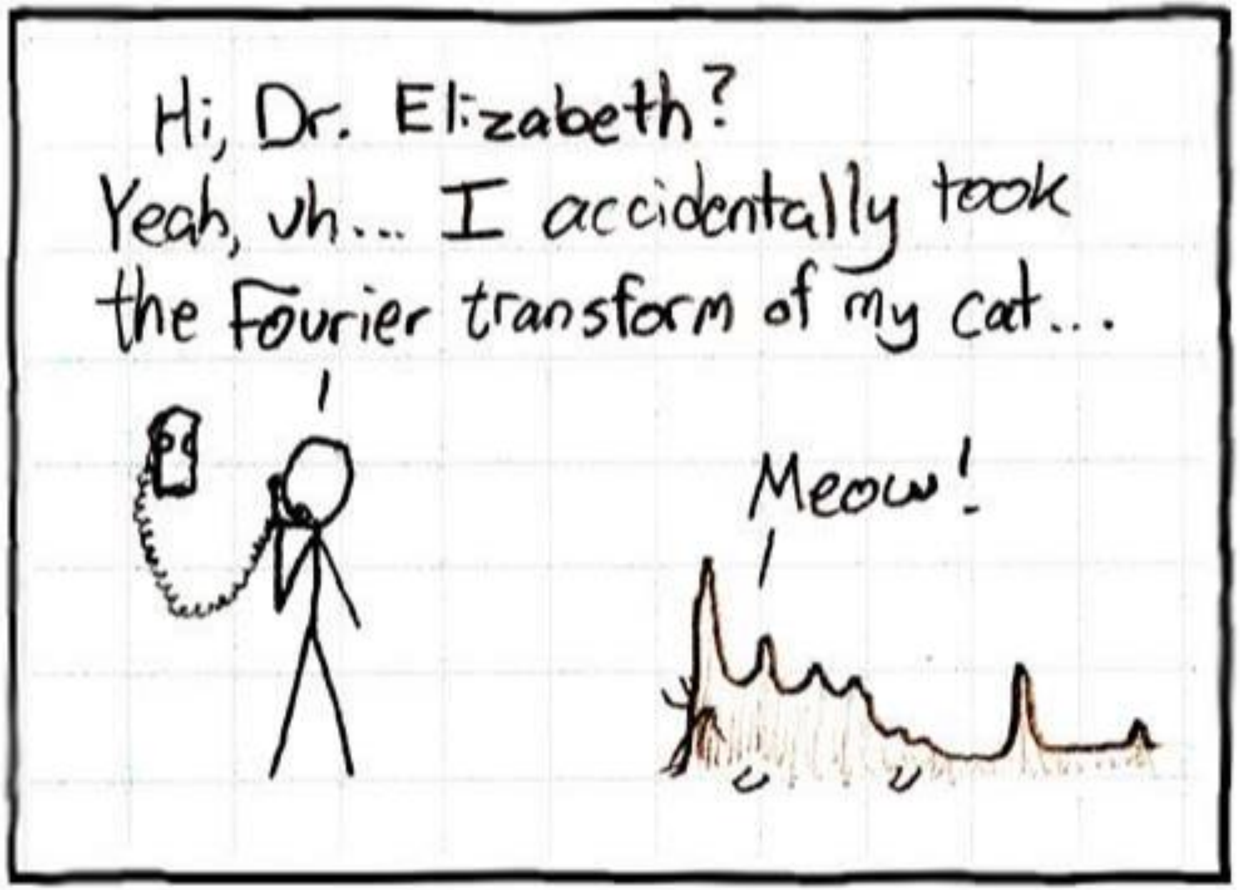


Rule 5: Multiple effects can be combined.





Rule 6: Duality


Hi, Dr. Elizabeth?
Yeah, uh... I accidentally took the Fourier transform of my cat...
Meow!

Questions?? Thoughts??

Lec 9 spectra

① — Extensive slideset for FT of Deterministic case.

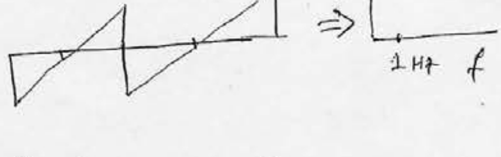
② — why use freq domain?

Recap

- — most physical phenomena are periodic (show figures)
- — gives further insights into a system (e.g. TF of a filter)
- — Makes analysis easier.

— What is a FT?

- — Is it just frequencies (periodicities)?
  - — If so, then this saw tooth should have a peak at 1 Hz

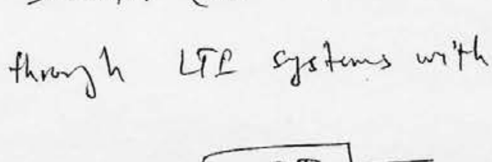


  - — but actually its FT has peaks at f, 2f, 3f, 4f, ...
- — that's because FT gives frequencies of sinusoids.

— why sinusoids?

- — They are smooth (derivatives easier) (Infinitely differentiable & integrable)
- — They pass through LTI systems with scaling & phase shift only

(Sinusoids eigen function of LTI)



- — Many natural vibrations are sinusoidal.

$Ax = \lambda x$

say A has two Evecs x_1, x_2

then $A\underline{u} = ?$

write $\underline{u} = \alpha_1 \underline{x}_1 + \alpha_2 \underline{x}_2$

computation simplified!

(show decomposition figure)

"bases" ← [essentially this is what FT does with sinusoids as eigenvectors]

— FT Definition.

$$FT \rightarrow \quad X(f) \triangleq \int_{-\infty}^{\infty} x(t)\, e^{-2\pi i f t}\, dt$$

$$IFT \rightarrow \quad x(t) = \int_{-\infty}^{\infty} X(f)\, e^{2\pi i f t}\, df \quad \leftarrow \text{more instructive}$$

$e^{2\pi i f t}$ ← complex sinusoid of freq. f

$X(f)$ → contribution of this sinusoid to the signal

— why use complex sinusoids?

$$e^{i\theta} = \cos\theta + i\sin\theta$$

→ Makes trig operations & analysis easier.

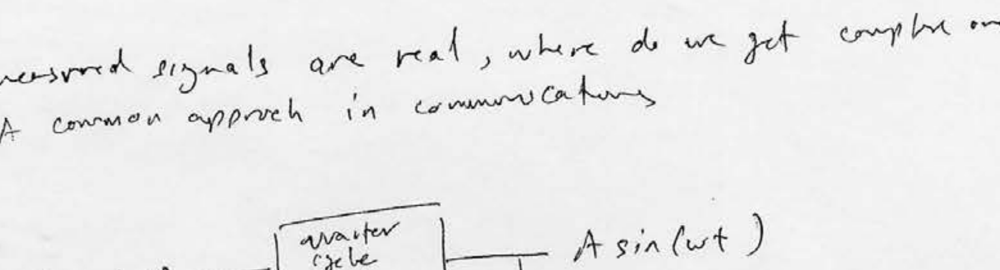
– e.g. how do you prove that

$$\cos x \cdot \cos y = \frac{1}{2}\left(\cos(x+y) + \cos(x-y)\right) \quad \left(\begin{smallmatrix}\text{both}\\ \text{sides}\\ \text{real}\end{smallmatrix}\right)$$

– easy if you use

$$\cos x = \frac{e^{ix} + e^{-ix}}{2} \quad \left(\begin{smallmatrix}\text{complex}\\ \text{only an intermediary tool}\end{smallmatrix}\right)$$

→ If all measured signals are real, where do we get complex ones from? A common approach in communications



— now play with $Ae^{i\omega t}$ and extract the real part out of the processed signal.

(3) — What if the signal/sequence is <u>stochastic</u>?

Show $A\cos(\omega t+\varphi)$ ~~+ sin...~~ figure and show amplitudes vary between realization ↓ could give misleading results about relative freq. strength.

— Problem: It would be unwise to take FT of just one realization.

— Soln: We know that $r_x(k)$ is deterministic and contains the same frequencies as the original signal. So we can study frequency content of $r_x(k)$ instead.

– e.g. as shown in sec ~~2.3.4~~ 2.4 of the book [Lindgren], for

$$X_t = A_0 + \sum_{k=1}^{n} A_k \cos(2\pi f_k t + \varphi_k) \quad . \quad A_k, \varphi_k \text{ indep. r.v.}$$

$$\varphi_k \sim U[0, 2\pi]$$

we have $r(\tau) = \sigma_0^2 + \sum_{k=1}^{n} \sigma_k^2 \cos(2\pi f_k \tau)$

→ FD counterpart of CVF for WSS processes.

(3) (29)

— PSD (Power Spectral Density)

— obtained by taking FT of $r_x(\tau)$: $\left\{ \begin{array}{l} R(f) = \int_{-\infty}^{\infty} r(\tau)\, e^{-2\pi i f \tau}\, d\tau \quad — (*) \\ r(\tau) = \int_{-\infty}^{\infty} R(f)\, e^{2\pi i f \tau}\, df \end{array} \right.$

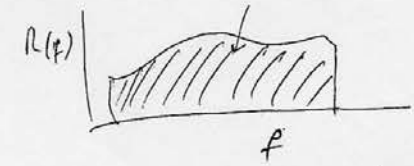
— tells us how the "mean effect" (power $(\frac{1}{2}E[A^2])$) of the process is distributed among various frequencies.

[condition for (*) LHS to exist] — $r(\tau)$ should be continuous and absolutely integrable, i.e.,

$$\int |r(\tau)|\, d\tau < \infty$$

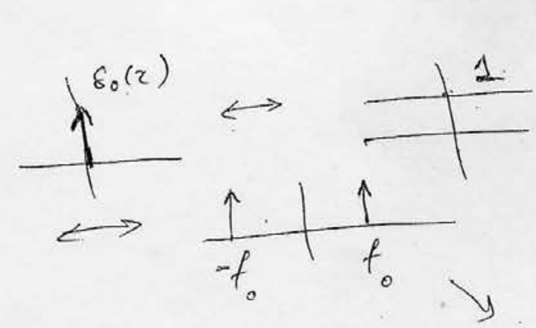
— Why call it "Power" spectral Density?

$$r(0) = \int_{-\infty}^{\infty} R(f)\, df$$



but $r(0) = V\{X(t)\}$ = Power of the signal (eg $\hat{r}(0) = \frac{1}{N}\sum_{t=1}^{N} x_t^2$ ← for zero-mean process)

— Use FT pairs in the book ~~(fig 2.19)~~



- unit Impulse (uncorrelated data) → $\delta_0(\tau)$ ↔ 1
- [PSD's of] Pure sinusoids → ↔ spikes at $-f_0$, f_0 → See next page (26)
- damped sinusoids → show figure. → damping increases line width. → Freq sets line location.
- Sinc function → show figure. Also, see ~~next page~~, page (28).
- Exponential decay. ↓ damping on DC case.

[Covered in Slides Already]

— Smooth $r(\tau)$ require less frequencies.

— PSD Properties

(i) $R(f)$ is symmetric, i.e, $R(f) = R(-f)$ (for $X \in \mathbb{R}$) ↓ real-valued RV.

(ii) ~~Positive~~ non-negative, i.e, $R(f) \geq 0 \quad \forall f$

(iii) Integrable

Defn — Continuous vs. Discrete spectra.

- Continuous ↓ when $R(f)$ is a continuous function (except for jump discontinuities)
- Discrete ↓ when $R(f)$ can be written as a sum of delta functions, i.e, $R(f) = \sum_k b_k \delta_{f_k}(f)$, $k = 0, \pm 1, \pm 2, \dots$

$\int_{-\infty}^{\infty} |r(\tau)| \, d\tau < \infty$

Helpful Theorems

— Th 4.3 — When is a PSD continuous?

(a) The spectrum of a continuous and absolutely integrable $r(\tau)$ is continuous.

— When is $r(\tau)$ a valid CVF?

(b) A continuous and absolutely integrable function $r(\tau)$ is a [valid] covariance function if its FT is symmetric, non-negative and integrable.

(unbiased estimator, $\hat{r}(\tau)$, does not satisfy this)!

- Theorem 4.1

- Existence of $R(f)$ for a stationary process with cvf $r(\tau)$

  : If the covariance function $r(\tau)$ of a stationary process $\{X(t), t \in \mathbb{R}\}$ is continuous then there exists a positive, symmetric, and integrable function $R(f)$ such that

$$r(\tau) = \int_{-\infty}^{\infty} e^{i2\pi f\tau} R(f)\, df$$ "

- so the only thing required is for $r(\tau)$ to be continuous!!

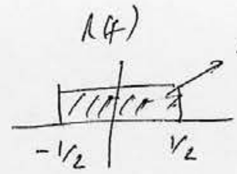
- Theorem 4.2 (Converse of Theorem 4.1)

- When does the IFT of a function $R(f)$ represent the cvf $(r(\tau))$ of a stationary process?

  : If $R(f)$ is a non-negative, symmetric, and integrable function (possibly containing delta functions) then its IFT defined as $r(\tau) = \int_{-\infty}^{\infty} e^{i2\pi f\tau} R(f)\, df$ represents the cvf of a stationary process.

Examples

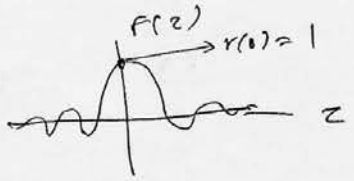
(1) $R(f) = \begin{cases} 1 & -\frac{1}{2} \le f \le \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$



1 at $\tau=0$ (using L'Hopital)

$$\Rightarrow r(\tau) = \int R(f) e^{2\pi i f \tau} df = \int_{-1/2}^{1/2} e^{2\pi i f \tau} df = \frac{e^{i\pi\tau} - e^{-i\pi\tau}}{2i\pi\tau} = \frac{\sin \pi\tau}{\pi\tau} \triangleq \text{sinc}(\tau)$$

$r(\tau)$ ↓ continuous.



(2) $r_x(\tau) = \sigma^2 \cos(2\pi f_0 \tau)$

~~$R(f) = \int_{-\infty}^{\infty} r(\tau) e^{-i2\pi f\tau} d\tau = \sigma^2 \int \cos(2\pi f_0 \tau) e^{-2i\pi f \tau} d\tau$~~

$$\Rightarrow r_x(\tau) = \frac{\sigma^2}{2}\left(e^{i2\pi f_0 \tau} + e^{-i2\pi f_0 \tau}\right)$$

$F\{1 \cdot e^{i2\pi f_0 \tau}\}$ using $F\{1\} = \delta_0(f)$ and $g(\tau)e^{i2\pi f_0 \tau} \leftrightarrow G(f-f_0)$

$$\Rightarrow F\{r_x(\tau)\} = R(f) = \frac{\sigma^2}{2}\left[\delta_0(f-f_0) + \delta_0(f+f_0)\right]$$

pg 219.

(3) Matching Figs.

Lec 5 DT signals, spectra, aliasing

— Recap

→ discussed CT-signal power spectra last time

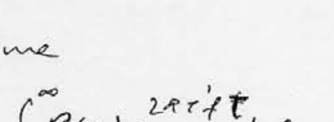
$$R(f) \triangleq \int_{-\infty}^{\infty} r(t)\, e^{-2\pi i f t}\, dt \iff r(t) = \int_{-\infty}^{\infty} R(f)\, e^{2\pi i f t}\, df$$

→ This lecture is about DT signals.

① Where do DT signals come from?

Naturally DT → # of students attending my lectures

→ can't have 1.5 lecture etc.



sampled versions of CT signals → Most common source of DT signals. e.g., recorded voice.

↓

Question arises → how should we sample to be able to faithfully reconstruct original CT signal?

→ discussed in "Aliasing".

② How do we define PSD of DT signals?

— consider a DT signal $\{X_t : t = 0, \pm 1, \pm 2, \dots\}$

with cov. fn. $r(\tau)$, τ discrete, we define

can be continuous or discrete (in freq-domain) ← $$R(f) \triangleq \sum_{\tau=-\infty}^{\infty} r(\tau)\, e^{-i2\pi f \tau}$$ valid only for $f \in [-\tfrac{1}{2}, \tfrac{1}{2}]$

→ $\tau \in \mathbb{Z}$ ← set of all integers.

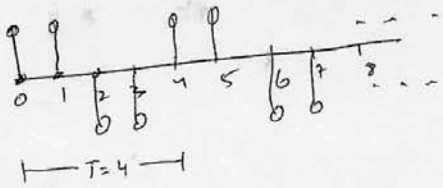
↓ First difference (Integral replaced by sum)

↓ second difference (frequency range of $r(\tau)$ limited)

always discrete by definition ← $$r(\tau) = \int_{-1/2}^{1/2} R(f)\, e^{i2\pi f \tau}\, df \quad \text{for } \tau \in \mathbb{Z}$$

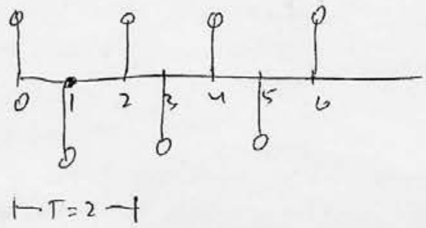
(2.1) why freq range is limited?

→ consider a DT signal of period $T = 4$ (i.e. $f = \frac{1}{T} = 1/4$)



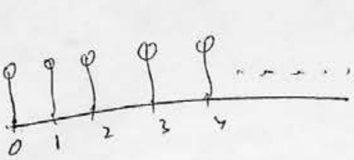
$(d = 1)$

→ consider a DT signal of period $T = 2$ (i.e. $f = \frac{1}{T} = 1/2$)



$(d = 1)$

→ can you plot a DT signal of period $T = 1$?



← indistinguishable from $T = \infty$ (i.e. $f = 0$, DC) $(d = 1)$
↓ never changing.

→ so $R(f)$ becomes invalid for $f \geq 1/2$

→ for sampling ~~time~~ distance d, freq range limited to

$$f \in \left(-\frac{1}{2d}, \frac{1}{2d}\right] \quad \left(\text{in standard notation } d = 1\right)$$

→ you could get $R(f)$ valid for higher freqs by choosing smaller d.

Example → white noise (DT)

$$r(\tau) = \begin{cases} \sigma^2 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases} \Rightarrow R(f) = \begin{cases} \sigma^2 & -\frac{1}{2} < f \leq \frac{1}{2} \\ 0 & \text{otherwise} \end{cases}$$

(3) How does sampling affect statistics of a r.p.?

— consider a stationary process

$$\{Y(t),\ t \in \mathbb{R}\}$$

with stat ⇒ m_y, $r_y(\tau)$, $R_y(f)$.

— let Z_t be generated by observing $Y(t)$ at times

$$t = 0, \pm d, \pm 2d, \ldots$$

— Then $Z_t = Y(t)$ at $t = 0, \pm d, \pm 2d \ldots$

(a) $m_z = E\{Z_t\} = E\{Y(t)\}\Big|_{t=0,\pm d,\pm 2d\ldots} = m_y$ (since $Y(t)$ is stationary.)

(b) $r_z(\tau) = C[Z_t, Z_{t+\tau}] = C[Y(t), Y(t+\tau)]\Big|_{t=0,\pm d,\pm 2d\ldots} = r_y(\tau)$ $\left[\text{for } \tau = 0, \pm d, \pm 2d \ldots\right]$ (since for stationary $Y(t)$, r_y depends on lag only.)

$r_z(\tau)$ defined only for $\tau = 0, \pm d, \pm 2d$.

(c) $R_z(f) = \sum_{k=-\infty}^{\infty} R_y(f + k f_s)$ for $-\frac{1}{2d} < f \leq \frac{1}{2d}$, where $f_s = \frac{1}{d}$ and $\frac{1}{2d} \triangleq f_n$ (Nyquist freq.)

(A) $r_z(\tau) = \int_{-f_n+0}^{f_n} e^{i2\pi f \tau} R_z(f)\, df$

Note: For existence of $R_z(f)$ and for checking validity of a DT CVF $r_z(\tau)$ see Theorem 4.4 (which is similar to the continuous case covered in Theorems 4.1 and 4.2)

(10) ~~(30)~~

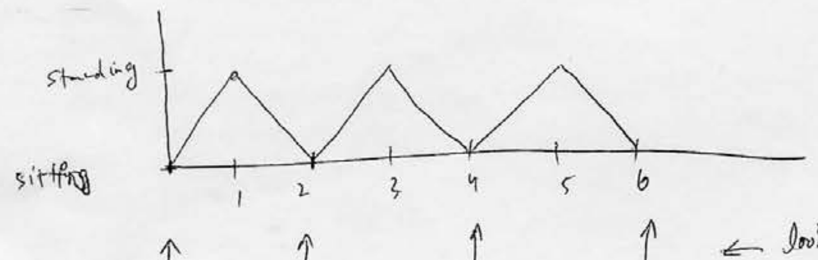
(4) Aliasing – by any other name ?

literature — Aliasing : a person assuming another identity . (Pseudonym)

SSP — Aliasing : A high frequency wrongly appearing as a low frequency due to improper sampling .

— when sampling, we want to capture the dynamics of a process (to get a true sense of reality).

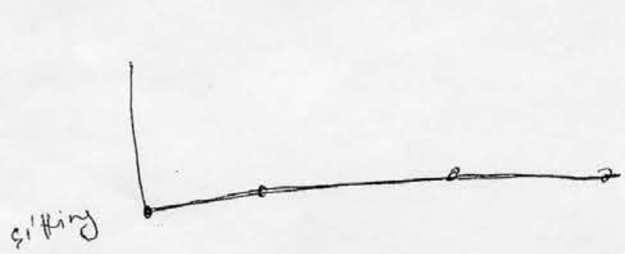
e.g — a student doing sit-ups every 2 second



his frequency of $\frac{1}{2}$ Hz appears as DC to me .

⇓

← looking back from board every 2 seconds, I see the student always sitting .



← $\frac{1}{2}$ Hz appearing as 0 Hz due to slow sampling.

→ I should look back more often!
→ Also show sampling figures.
→ But exactly how often ?
→ Aliasing formula gives us the exact relation.

11

$$R_z(f) = \sum_{k=-\infty}^{\infty} R_y(f+kf_s) \quad \text{for} \quad -\frac{1}{2d} < f \le \frac{1}{2d}$$

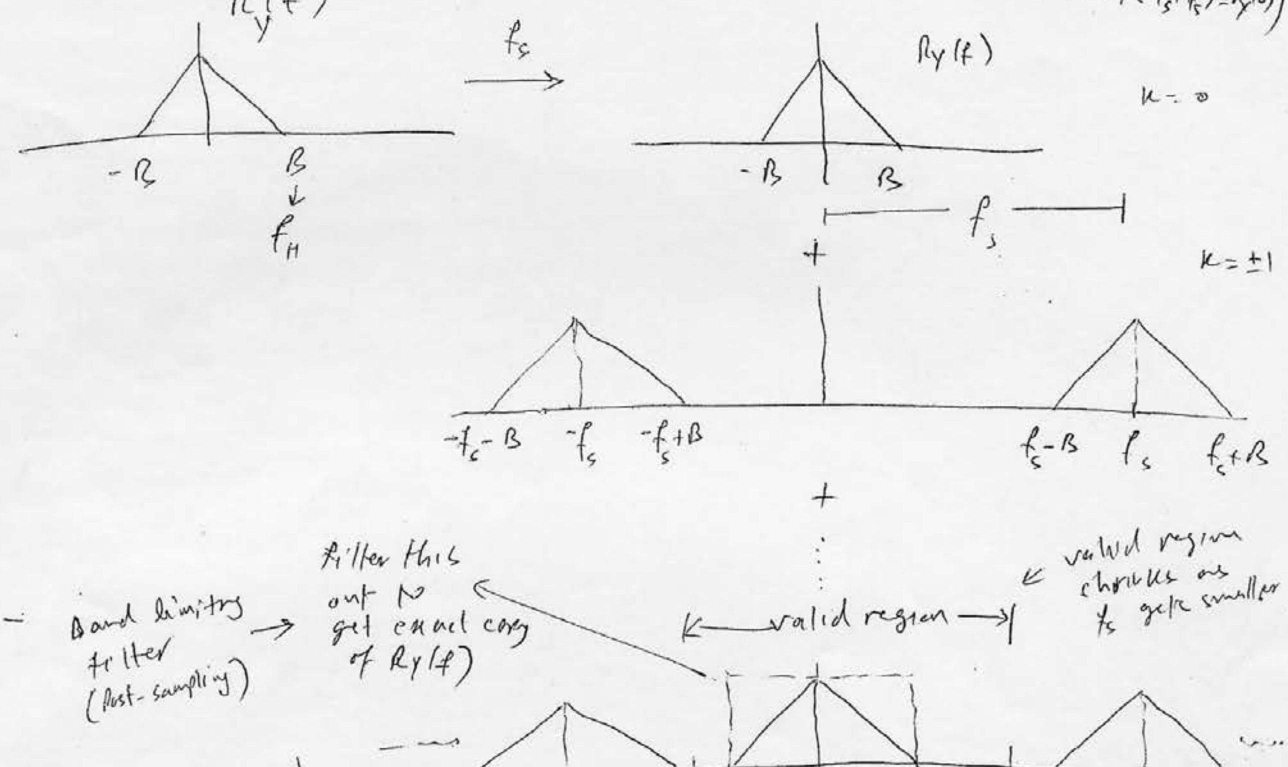
$$= R_y(f) + R_y(f \pm f_s) + R_y(f \pm 2f_s) + \dots$$

$R_y(f)$ → original CT signal spectrum

the remaining terms → shifted copies of $R_y(f)$

(e.g. for $R_y(f)$ having single peak at 0 Hz, $R_y(f+f_s)$ would be zero at $R_y(f_s)$ but would have the peak at $-f_s$ since $R_y(-f_s+f_s) = R_y(0)$)

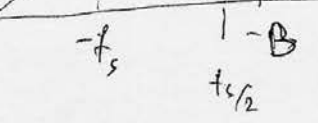
→ so for bandlimited signal, e.g.



→ Aliasing occurs when some of the replicas enter the main region of interest

→ from the figure its clear that to avoid this we need

$$f_s > 2f_H$$

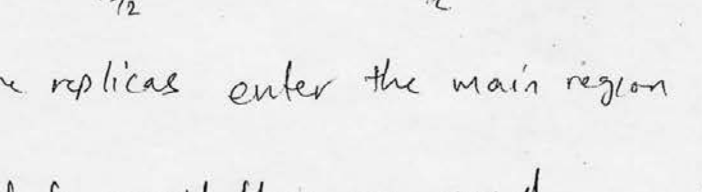


how do we satisfy

Q: → for band-unlimited signal $f_s > 2f_H$?

soln → Define region of interest and filter in that part before sampling.

(so) ⇒



(5) Shannon's Sampling Theorem

– we previously saw what happens to a CT process (and its PSD) when we sample it at some frequency f_s, and there we came across the notion of Nyquist frequency $f_n = 2 f_s$.

– Next, we look at the question "given uniformly sampled discrete values of a stationary process, when can we perfectly reconstruct the original continuous process"?

– Th. 4.7 Shannon's Sampling Theorem

"If the stationary process is band-limited to $[-f_0, f_0]$, then it can be perfectly reconstructed from its ~~et~~ values at discrete time points, spaced $t_0 = \frac{1}{2f_0}$ apart (i.e. $f_s = 2f_0$)".

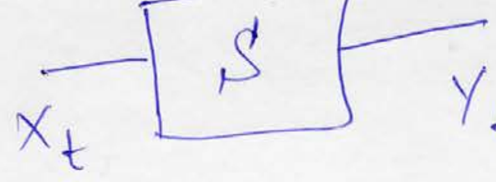
→ Note that the above formulation applies to uniform sampling.

Lecture 10 Filters

— Anything that operates on a process (signal)

— can be a mathematical function or something physical or an algorithm.

— we often find it beneficial (or convenient) to model filters in terms of their inputs and outputs.



— S can then come from a variety of fields (mechanics, electronics, biological.....)

— of course, if a complete physical description (model) of S is possible (e.g from first principle) we could go for that.

— But that is not always convenient or possible

— In fact, often we only have access to input and output behaviour of S and then ask ourselves: how can we model

S based on these inputs and observed outputs?"

— Such modeling framework (called "black-box model") has the added advantage of being independent of the field of study.

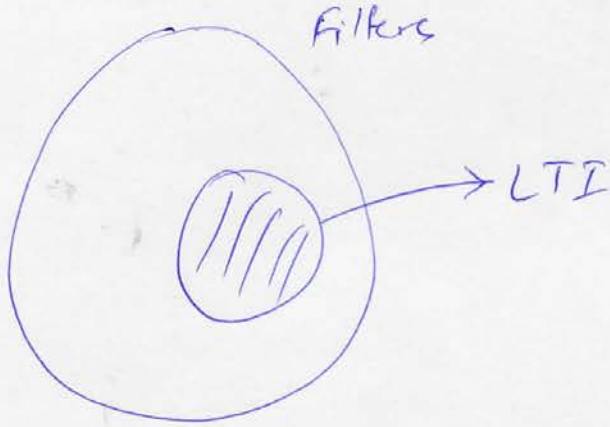
Question: How do we tell the "response" (output called) of a filter to an arbitrary input?

— make a database of outputs?
  — not very effective in general.

— Is it even possible to have a general formula?

Divide & Rule: One strategy is to look at a subclass of filters and answer our question for that subclass.

— In fact, that is what we do, and the subclass is called Linear Time-Invariant (LTI) filters.



- This approach makes the problem more managoable.
- In fact, solutions developed for this subclass can be used for non-linear or time-varying (TV) filters by exploiting local linearity (and/or time-invariance

<u>But</u> what about the "arbitrary" input?

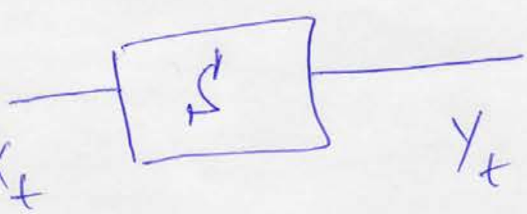
- can we have a generic model for that?
- In fact we can, and for that we will make use of a special function called the "Delta" function.

<u>So</u> Today we will see

ⓐ how to define the LTI class
ⓑ How to use Delta function to model processes (and <u>what</u> exactly is the Delta function)

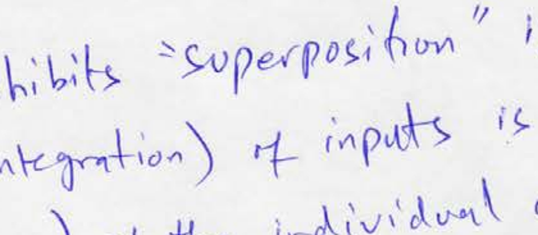
(c) How to develop a "black-box" model of an LTI filter using convolution and Impulse Response (and what are "convolution" and "Impulse Response" any way?)

(a) Defining LTI



$$Y_t \triangleq S\{X_t\}$$

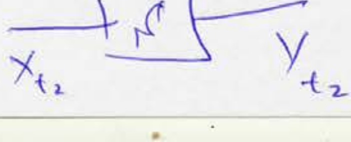
— A filter is "Homogeneous" if scaling at the input leads to the same scaling at the output, i.e.



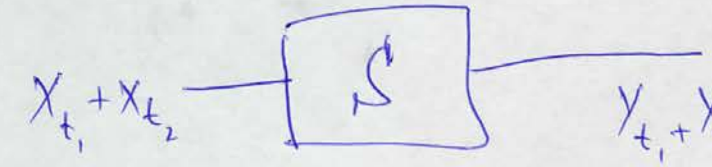
$$S\{\alpha X_t\} \stackrel{?}{=} \alpha S\{X_t\}$$

— A filter exhibits "superposition" if the response to sum (or integration) of inputs is simply the sum (or integration) of the individual outputs, i.e., say

and

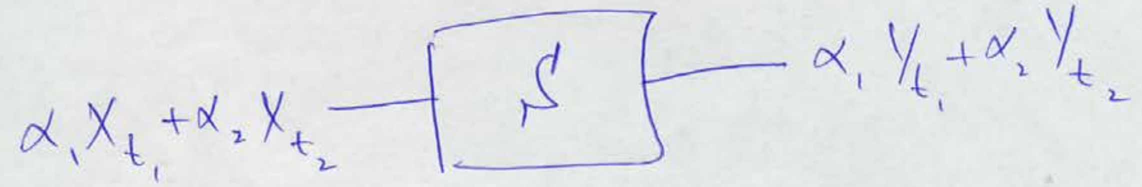


Then, for a filter with superposition property we have



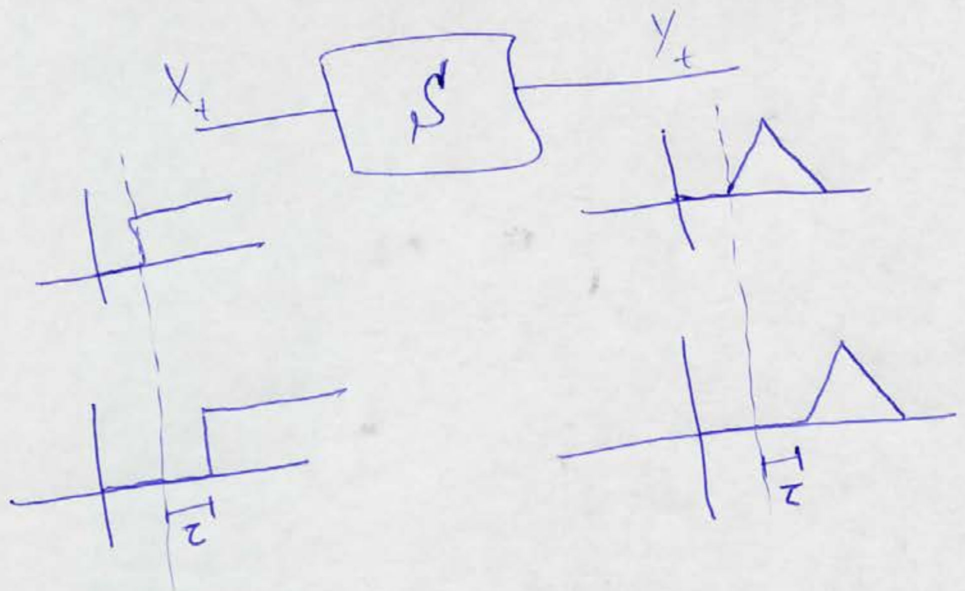
$$S\{x_{t_1} + x_{t_2}\} \stackrel{?}{=} S\{x_{t_1}\} + S\{x_{t_2}\}$$

— A filter is "linear" if it is both <u>homogeneous</u> and exhibits <u>superposition</u>, i.e., for a Linear system we have



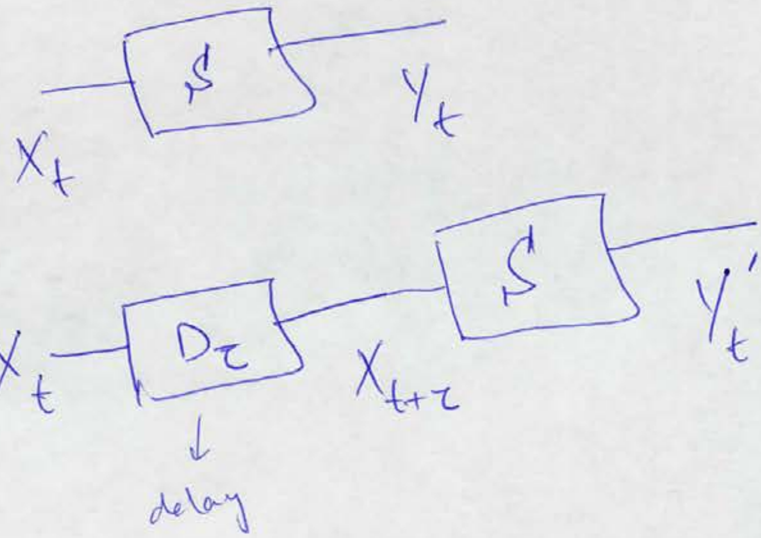
Linearity: $$S\{\alpha_1 x_{t_1} + \alpha_2 x_{t_2}\} = \alpha_1 S\{x_{t_1}\} + \alpha_2 S\{x_{t_2}\}$$

— In addition, a filter is Time-Invariant (TI) if its response (behavior) does not depend on "when" the input is given, instead, ~~the~~ any delay (or advance) in the input results in an equivalent delay (or advance) in the output with no other change in the output.



so, say

and



is $Y'_t = Y_{t+\tau}$? (i.e. does a shift in input result only in an equal shift in output and no other effects?)

if so, then S is TI.

— so, for a TI filter S we have

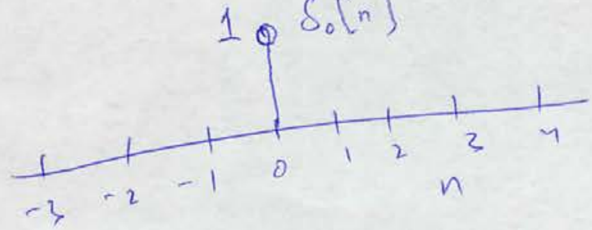
$$S\{X_{t+\tau}\} \triangleq S\{D_\tau(X_t)\} \stackrel{?}{=} D_\tau\{S\{X_t\}\}$$

(b) What is the Delta function and how it can represent an arbitrary process?

(i) Discrete-Time Case:

— A DT Delta function is just a unit sample, i.e.,

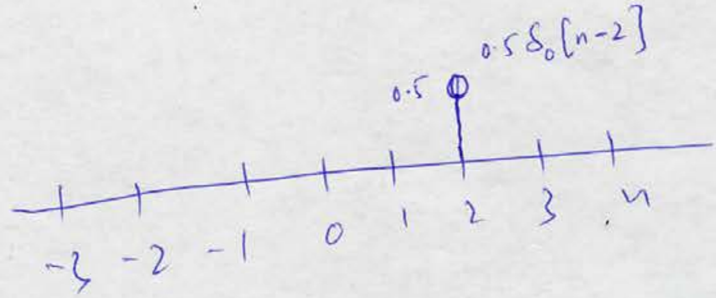
$$\delta_0[n] = \begin{cases} 1 & n = 0 \\ 0 & o/w \end{cases}$$



— It can be scaled and shifted

$$\alpha\,\delta_0[n-k] = \begin{cases} \alpha & n = k \\ 0 & n \neq k \end{cases}$$

e.g. for $k = 2$ and $\alpha = 0.5$ we have



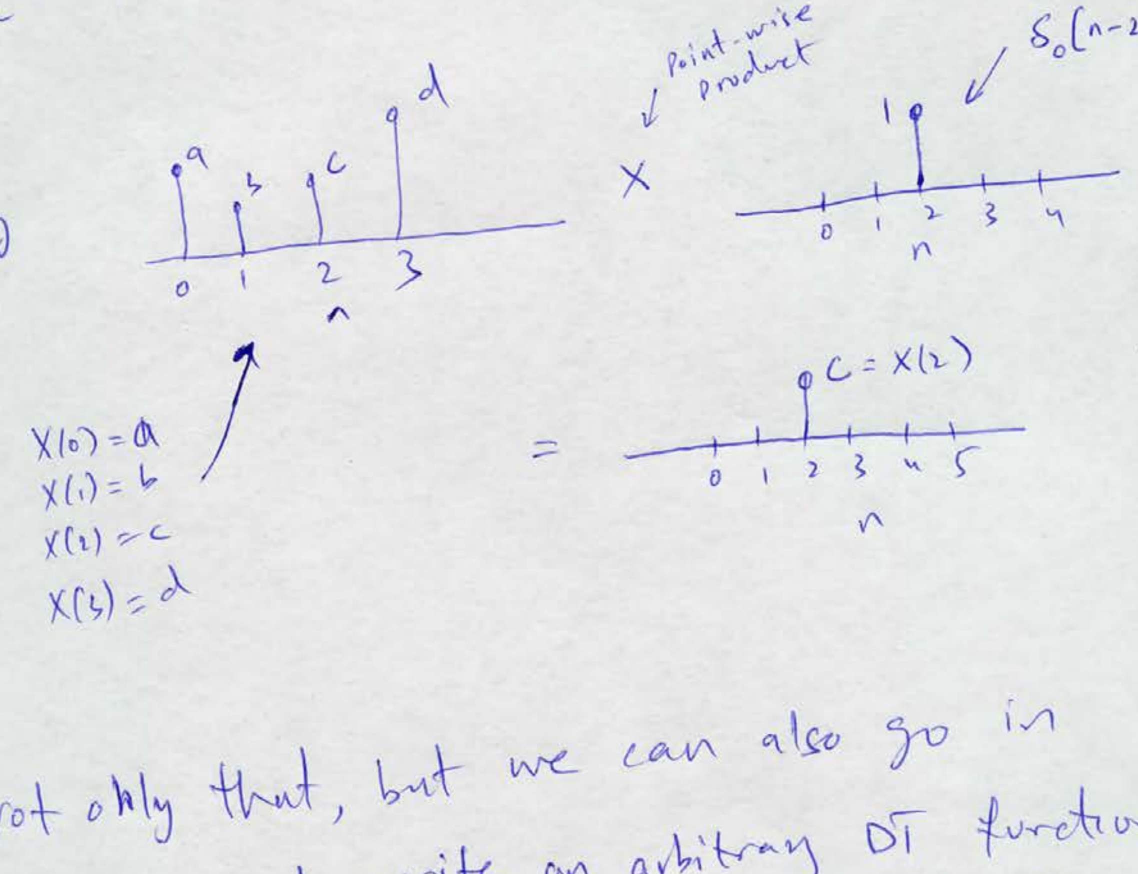
— we also call it an "ideal sampler" for DT functions (helps extract a single value, i.e. sample, from a function), e.g., consider a DT function $\{x[n]\}$

then we can represent the kth value of this function by

$$x[k] = \sum_{n=-\infty}^{\infty} x[n]\,\delta_0[n-k]$$

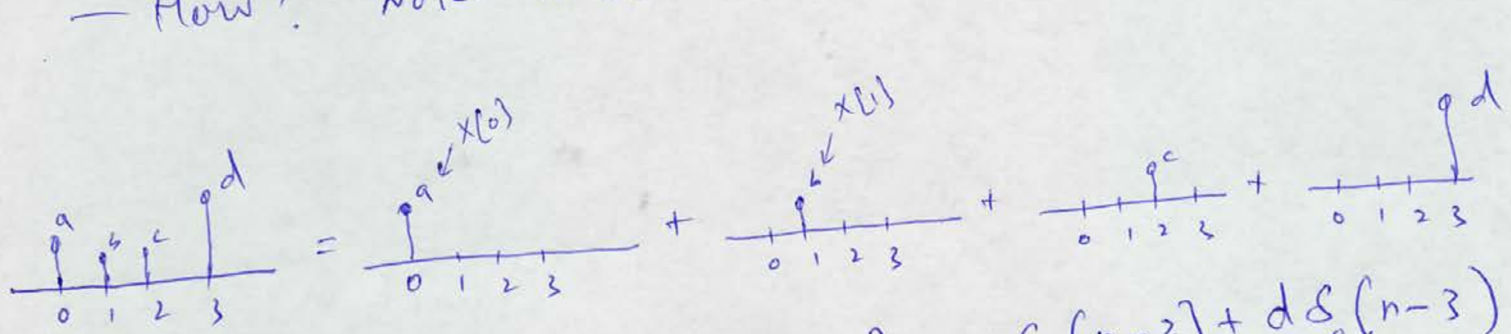
(how? expand RHS and see!)

— e.g.



— not only that, but we can also go in reverse and write an arbitrary DT function as a "sum of scaled and shifted Delta functions"

— How? note that



$$x[n] = a\delta_0[n] + b\delta_0[n-1] + c\delta_0[n-2] + d\delta_0[n-3]$$

sum of scaled and shifted Delta functions

In general

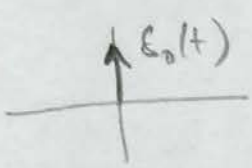
$$x[n] = \sum_{k=-\infty}^{\infty} x(k)\,\delta_0[n-k]$$

(ii) Continuous-Time Case

— For CT case, we define Delta function by its properties

ⓐ $\delta_0(t) = 0 \quad t \neq 0$ → zero everywhere except at zero

and ⓑ $\int_{-\infty}^{\infty} \delta_0(t)\,dt = 1$ → has total area 1.



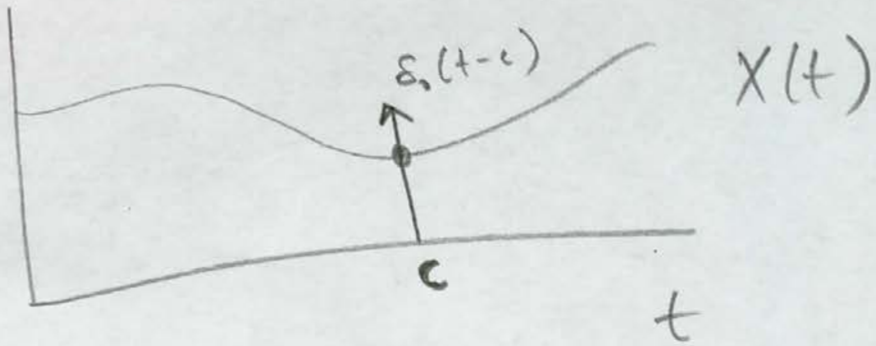
— we also sometimes define it by its "Sampling property"

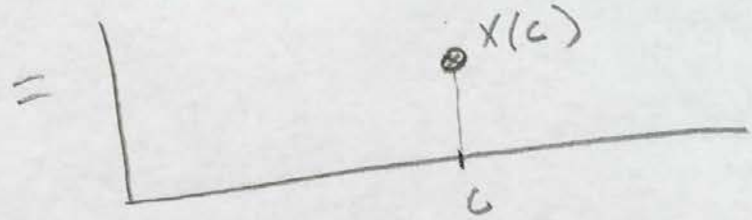
$$\int_a^b f(t)\,\delta_0(t-c)\,dt = \begin{cases} f(c) & \text{if } c \in (a,b) \\ 0, & \text{otherwise} \end{cases}$$

— Thus we can also use it as an ideal sampler of CT functions, as

$$X(c) = \int_{-\infty}^{\infty} X(t)\,\delta_0(t-c)\,dt$$

(how? By definition of the CT Delta Function!)





— And just as in the DT case, we can ~~use~~ write arbitrary CT functions as a "integral of scaled and shifted impulses"

$$X(t) = \int_{-\infty}^{\infty} X(u)\,\delta_0(t-u)\,du$$

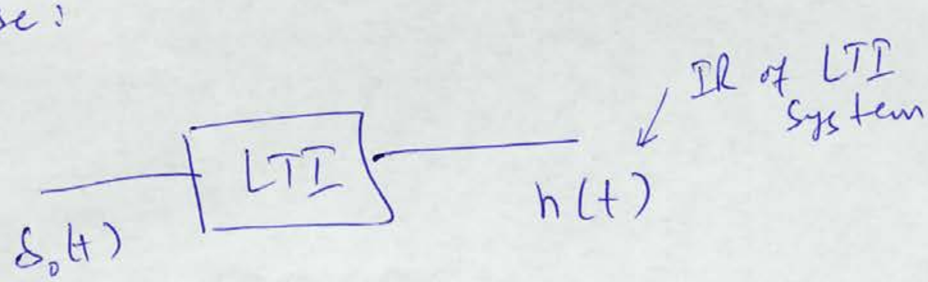
— Finally, we remark that Delta Function is also called "IMPULSE"

(c) How do delta & LTI come together to facilitate black-box modeling of filters?

— related vocabulary < Convolution / Impulse Response.

— Impulse Response (IR): The response of a filter to a Delta input ("= Impulse")

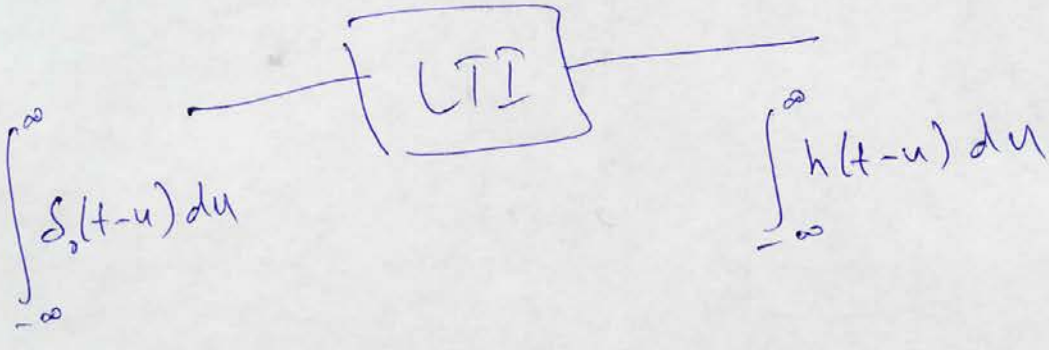
— LTI case:



— now, since the System is LTI, any scaling of input should gives same scaling in output

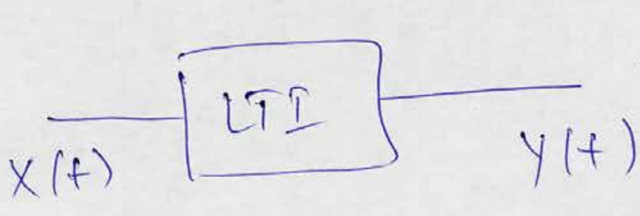
— Any time-shift in input should result only in an equivalent shift in output

— And any superposition of inputs, should result in superposition of outputs related to the individual inputs



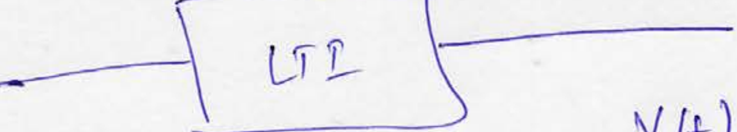
— Now suppose we have that
- — System/filter is LTI
- — Input given is $X(t)$
- — IR of the filter is known, $h(t)$

— and we ask
- — what will be the output $Y(t)$?



— Utilizing something we saw before (that an arbitrary process can be written as an integral [CT-case] or sum (DT-case) of scaled and shifted impulses (Delta functions) and exploiting the properties of LTI, we conclude that the output will be an integral/sum of similarly scaled and shifted Impulse responses!

i.e



$$x(t) = \int_{-\infty}^{\infty} x(u)\, \delta_0(t-u)\, du$$

$$y(t) = \int_{-\infty}^{\infty} x(u)\, h(t-u)\, du$$

This kind of integral, in fact, has a name (or "operator") dedicated to it called "Convolution"

Defn. convolution $\left(\begin{array}{l} \text{operator} \\ \text{symbol} \end{array} : * \right)$

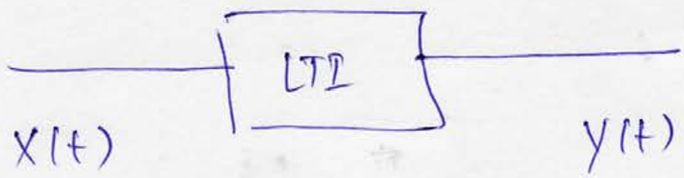
$$DT: \quad f(n) * g(n) \triangleq \sum_{k=-\infty}^{\infty} f(k)\, g(n-k)$$

$$CT: \quad f(t) * g(t) \triangleq \int_{-\infty}^{\infty} f(\tau)\, g(t-\tau)\, d\tau$$

(see properties of convolution, such as commutative etc., in some basic text)

— so, we can say that for an LTI filter, the output to an input $x(t)$ is the convolution b/w $x(t)$ and the filter's IR $h(t)$

i.e



$$y(t) = \int_{-\infty}^{\infty} x(u)\, h(t-u)\, du$$

$$= x(t) * h(t)$$

— The beauty of LTI is that it is fully characterized by its impulse response h_t.

— The FT of h_t is called "TF". (Transfer Function)

$$H(f) = \int_{-\infty}^{\infty} h(t)\, e^{-2\pi i f t}\, dt \iff h(t) = \int_{-\infty}^{\infty} H(f)\, e^{i 2\pi f t}\, df$$

— Other ~~desirable~~ properties of interest (for [LTI] filters)

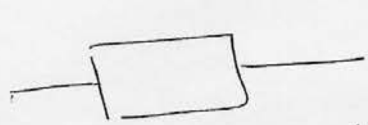
(i) causality: output ~~should~~ when does not depend on future inputs.

— e.g. for LTI systems

$$y(t) = \int_{-\infty}^{\infty} x(u)\, h(t-u)\, du = \int_{-\infty}^{\infty} x(t-u)\, h(u)\, du$$

∴ For LTI causality $h(t) = 0$ for $t < 0$ (e.g. check above)

(ii) BIBO stability: Bounded input ~~should~~ when gives bounded output



$|x(t)| \le A < \infty$ $\qquad |y(t)| \le B < \infty$

Given $|x(t)| \le A < \infty$ → For LTI

$$|y(t)| = \left| \int_{-\infty}^{\infty} h(u)\, x(t-u)\, du \right|$$

$$\le \int_{-\infty}^{\infty} |h(u)|\, |x(t-u)|\, du$$

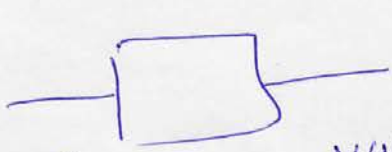
Cauchy-Schwartz inequality.

$$\le A \int_{-\infty}^{\infty} |h(u)|\, du$$

stable if $\int_{-\infty}^{\infty} |h(u)|\, du < \infty$ ← i.e. IR should be absolutely integrable.

(iii) Instantenous vs. Dynamic : If the system output depends only on the current input (and not on future or past inputs) it is said to be an Instantaneous System (also known as "Memoryless) otherwise it is called Dynamic (or "Memory System")

e.g



$y(t) = 4x(t) + x^2(t)$ ← Instantaneous

$y(t) = 4x(t) + x^2(t+1)$ ← Dynamic

— For LTI systems we have

$$y(t) = \int_{-\infty}^{\infty} x(t-u)\,h(u)\,du$$

∴ For LTI systems being instantaneous means

$$h(t) = \alpha\,\delta_0(t)$$ (i.e IR must be an impulse) (at $t=0$ with a possible scaling α)

— otherwise system is Dynamic (depends on past and/or future inputs also).

Lecture 11

Recap Last time we saw filters and their subclass LTI filters.

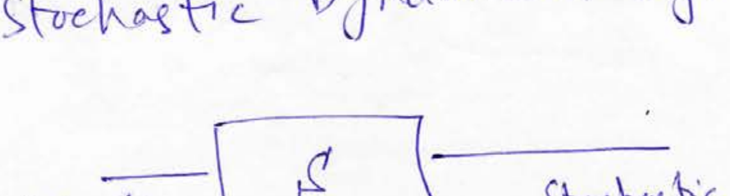
- we developed a black-box model for the LTI fillters in terms of convolution (b/w input & IR)
- we saw various properties of filters & LTI filters
- What we did not talk about was whether the input was deterministic or stochastic.

Today We will specifically consider the case where the input is a stochastic process

- and we will mostly limit ourselves to stationary stochastic processes and LTI filters.
- we will also see how to define the operations (such as integration involved in convolution for LTI filter output) for stochastic processes

  i.e some stochastic Calculus !!

(1) (a) **Defn**: A "Stochastic Dynamical System"



= A system/filter where input and output are stochastic processes"

(b) Important Assumption → In our work, we will assume that the system itself does not add any randomness.

This implies that → same input always gives the same output

otherwise → things would get a lot more complicated (⇒ Advanced texts)

(2) What a Stochastic Dynamical System Does to a random process input (and its statistics)

(a) Defn. 6.1 In a CT linear & TI filter the output stochastic Process $\{Y(t)\}$ is obtained from the input stochastic process $\{X(t)\}$ through convolution with a generalized function $h(t)$ (called the "Impulse Response") as

$$Y(t) = \int_{-\infty}^{\infty} h(t-u)\, X(u)\, du = \int_{-\infty}^{\infty} h(u)\, X(t-u)\, du$$

Note: A "generalized" function is a function that relaxes the strict definitions of a function (such as Delta function or a regular functions containing Deltas etc.)

(b) How stable LTI affects statistics of a WSS process

Th. 6.1 "If input to LTI filter is stationary then so is the output"

WSS → [LTI] → WSS

mean

$$m_y = E\{y(t)\} = E\left[\int_{-\infty}^{\infty} h(u)\, x(t-u)\, du\right]$$

$$= \int_{-\infty}^{\infty} h(u)\, E\{x(t-u)\}\, du = m_x \int_{-\infty}^{\infty} h(u)\, du$$

by Th. 5.1 and stability.

(see outline of the proof in 6.3.3 [Lindgren])

cvf

$$r_y(\tau) = C\{y(t), y(t+\tau)\} = C\left[\int_{-\infty}^{\infty} h(u)\, x(t-u)\, du, \int_{-\infty}^{\infty} h(v)\, x(t+\tau-v)\, dv\right]$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(u)\, h(v)\, \underbrace{r_x(\tau+u-v)}\, du\, dv$$

by th. 5.1 and stability

PSD

Use $r_x(\tau) = \int_{-\infty}^{\infty} R_x(f)\, e^{i2\pi f\tau}\, df$ in

$$\Rightarrow r_y(\tau) = \iint h(u)\, h(v) \int e^{i2\pi f(\tau+u-v)} R_x(f)\, df\, du\, dv$$

$$= \int e^{i2\pi f\tau}\Big(\underbrace{\int h(u)\, e^{i2\pi f u}\, du}_{\bar{H}(f)}\ \underbrace{\int h(v)\, e^{-i2\pi f v}\, dv}_{H(f)}\Big) R_x(f)\, df$$

$$= \int e^{i2\pi f\tau} |H(f)|^2 R_x(f)\, df \quad \text{—— (A)}$$

But $r_y(\tau) = \int e^{i2\pi f\tau} R_y(f)\, df$ —— (B) ← by defn.

comparing (A) & (B) $\Rightarrow$ $R_y(f) = |H(f)|^2 R_x(f)$

TF note that in the PSD we have used the definition

$$H(f) = \int_{-\infty}^{\infty} h(u) e^{-i2\pi f u} du$$

which is the Fourier Tx. of the IR. $H(f)$ in fact has a name = "Transfer Function".

Amplitude & Phase Response We may further expand $H(f)$ to see how the LTI filter affects the amplitudes and phases of the incoming frequencies (found in the input process). These will be called "Amplitude Response" and "Phase Response", respectively

$|H(f)|$ = Amplitude Response

$\angle H(f)$ = Phase Response

$$H(f) = |H(f)| e^{i \angle H(f)}$$

Interpretation If the input has a sinusoidal frequency component $[f_k, \alpha_k, \phi_k]$ (amplitude α_k, phase ϕ_k) then the output will have (or we say "the LTI filter will alter it to")

$$[f_k, \alpha_k |H(f)|, \phi_k + \angle H(f)]$$

5

(c) Since we have PSD in terms of TF, would be nice to have everything else in TF as well.

output $y(t) = h(t) * x(t)$

$$\Rightarrow Y(f) = H(f) X(f)$$ (convolution ↕ product)

mean $$m_y = m_x \int_{-\infty}^{\infty} h(u)\, du = m_x \underbrace{\int_{-\infty}^{\infty} h(u) e^{-i2\pi f 0} du}_{H(0)} = m_x H(0)$$

variance $$r_y(\tau) = \int e^{i2\pi f \tau} |H(f)|^2 R_x(f)\, df$$

$$\Rightarrow r_y(0) = \int |H(f)|^2 R_x(f)\, df$$

(d) Repeat all for DT

$$y(t) = \sum_{u=-\infty}^{\infty} h(u)\, x(t-u)$$

$$m_y = m_x \sum_{u=-\infty}^{\infty} h(u) = m_x H(0)$$

$$r_y(\tau) = \sum_{u=-\infty}^{\infty} \sum_{v=-\infty}^{\infty} h(u) h(v) r_x(\tau+u-v)$$

$f \in \left(-\frac{1}{2}, \frac{1}{2}\right] \leftarrow$ $$H(f) = \sum_{t=-\infty}^{\infty} h(t) e^{-i2\pi f t} \Longleftrightarrow h(t) = \int_{-1/2}^{1/2} H(f) e^{i2\pi f t} df$$

$$R_y(f) = |H(f)|^2 R_x(f)$$

(3) Stochastic Calculus < Convergence, Continuity, Differentiation, Integration

(a) Convergence

We say that a sequence of RVs $\{X_n\}$ converges to an RV X in the quadratic mean [q.m] sense (also called "mean-square" [m.s.] sense) iff

$$\lim_{n\to\infty} E\left[(X_n - X)^2\right] = 0$$

(b) Continuity of an RP.

(i) An RP $\{X(t)\}$ is called "continuous in quadratic mean" at time t if $X(t+h) \xrightarrow{q.m} X(t)$ as $h \to 0$

i.e. $E\left[\left(X(t+h) - X(t)\right)^2\right] \to 0$ as $h \to 0$

(ii) A stationary RP $\{X(t)\}$ is "everywhere continuous in q.m." if

$$E\left[\left(X(t+h) - X(t)\right)^2\right] = 2\left(r_X(0) - r_X(h)\right) \to 0 \text{ as } h\to 0$$

(i.e. if the covariance function is continuous for $t=0$)

→ ~~Stochastic calculus.~~

(c) → Differentiability of a r.p.

A stoch. process $\{X(t)\}$ is diffrentiable in quadratic mean (denoted q.m.) with derivative $\{X'(t)\}$ if

$$E\left\{\left[\frac{X(t+h)-X(t)}{h} - X'(t)\right]^2\right\} \to 0 \quad \text{as} \quad h \to 0 \quad \forall t$$

"mean" and "quadratic" included for r.p.

Recall: for deterministic $s(t)$

$$\lim_{h\to 0} \frac{s(t+h)-s(t)}{h} = \frac{ds(t)}{dt} = s'(t)$$

if exists

Denoted

$$\frac{X(t+h)-X(t)}{h} \xrightarrow{q.m.} X'(t) \quad \text{as } h \to 0$$

Th 6.3 → Differentiability of a weakly stationary process

Let $\{X(t)\}$ be a weakly stationary process with covariance function $r_x(\tau)$, then

(a) $\{X(t)\}$ is differentiable in q.m. iff $r_x(\tau)$ is twice differentiable for every τ. Further, the derivative $\{X'(t)\}$ is also weakly stationary with

$$m_{X'} = 0$$

$$r_{X'}(\tau) = -r_X''(\tau) \quad — (*)$$

$$R_{X'}(f) = (2\pi f)^2 R_X(f) \quad — (\Delta)$$

Proof in the book.

useful for (a) → (b) $r_x(\tau)$ is twice differentiable iff

$$\int_{-\infty}^{\infty} (2\pi f)^2 R_X(f)\, df < \infty$$

(c) The derivative of Gaussian is also Gaussian.

→ we show that for a differentiable WSS $X(t)$ $\to m_x, r_x(\tau)$
(~~5-10~~) and (~~5-11~~) hold.
(☆) (△)

(i) if $\frac{X(t+h) - X(t)}{h} \xrightarrow{q.m} X'(t)$ ↙ recall that this equals $X'(t)$ in q.m. sense.

then $$m_{X'} = E\left[\lim_{h\to 0} \frac{X(t+h) - X(t)}{h}\right] \quad \left(\begin{array}{c}\text{by} \\ \text{Theorem} \\ \text{A.4}\end{array}\right)$$

$$= \lim_{h\to 0}\left[E\, \frac{X(t+h) - X(t)}{h}\right] = \lim_{h\to 0} \frac{m_X - m_X}{h} = 0$$

(ii) $$r_{X'}(\tau) = C\left[\lim_{k\to 0} \frac{X(t+k) - X(t)}{k}, \lim_{h\to 0} \frac{X(t+\tau+h) - X(t+\tau)}{h}\right]$$

$$= \lim_{h\to 0} \lim_{k\to 0} E\left[\frac{X(t+k) - X(t)}{k}, \frac{X(t+\tau+h) - X(t+\tau)}{h}\right]$$

$$= \lim_{h\to 0} h^{-1} \lim_{k\to 0} \left\{\frac{r_X(\tau+h-k) - r_X(\tau+h)}{k} - \frac{r_X(\tau-k) - r_X(\tau)}{k}\right\}$$

$$= \lim_{h\to 0} \frac{-r_X'(\tau+h) + r_X'(\tau)}{h} = -r_X''(\tau)$$

(iii) $$r_X(\tau) \Longleftrightarrow R_X(f)$$

$$r_X'(\tau) \Longleftrightarrow (2\pi i f) R_X(f)$$

$$r_X''(\tau) \Longleftrightarrow (2\pi i f)^2 R_X(f)$$ (~~what about the negative sign?~~)

now, since $r_{X'}(\tau) = -r_X''(\tau)$, we have

$$r_{X'}(\tau) \Longleftrightarrow (2\pi f)^2 R_X(f) \doteq R_{X'}(f)$$

(d) ~~Th 6.5~~ → Cross-covariance of $\{X(t)\}$ and $\{X'(t)\}$

Th. 6.4

→ if $\{X(t)\}$ satisfies Th 6.3 then

$$r_{X,X'}(t,t+\tau) = r_X'(\tau)$$

$$\swarrow$$

$$C[X(t), X'(t+\tau)]$$

→ ~~thus also leads to~~ Further

$$r_{X,X'}(t,t) = r_X'(0) = 0$$

Since $r_X(\tau)$ is a symmetric function, its derivative at $\tau=0$ must be zero (e.g. draw some $r(\tau)$ functions and see)

i.e $X(t)$ and $X'(t)$ are uncorrelated when taken at the same time (but are usually uncorrelated if taken at different times)

(e) ~~Th 6.5~~ → Integration of a wss process. (sec. 6.3.3)

→ If $\{X(t)\}$ is a wss process with continuous covariance function, and $g(u)$ is a bounded and integrable function over $[a,b]$, then the integral

$$\int_a^b g(t)\, X(t)\, dt \text{ exists}$$

and one can change order of Expectation and integration

$$E\left\{ \int_a^b g(t)\, X(t)\, dt \right\} = \int_a^b g(t)\, E\{X(t)\}\, dt = m_X \int_a^b g(t)\, dt$$

Further,

$$C\left\{ \int_a^b g(s)\, X(s)\, ds,\ \int_c^d h(t)\, X(t)\, dt \right\} = \int_{s=a}^{b} \int_{t=c}^{d} g(s)\, h(t)\, r_X(s-t)\, dt\, ds.$$

(4) Cross-covariance & cross-spectrum.

need
- we've studied $r_X(s,t) = C[X(s), X(t)]$
- But often we want to study correlations between input and output. For that we need to define cross-cov and cross-spectra.

Defn. — $r_{X,Y}(s,t) = C[X(s), Y(t)]$ ← tells of linear dependence between different processes.

stationarity
— If $r_{X,Y}(s,t)$ depends only on $s-t$ then $\{X(t)\}$ and $\{Y(t)\}$ are "stationary correlated", and cross-cov is

$$r_{X,Y}(\tau) = C[X(t), Y(t+\tau)]$$

non-symmetric
— Notice that unlike $r_X(\tau)$, $r_{X,Y}(\tau)$ is not necessarily symmetric

$$r_{X,Y}(\tau) \neq r_{X,Y}(-\tau)$$

(and in general $r_{X,Y}(\tau) \neq r_{Y,X}(\tau)$ i.e., order matters)
(although $r_{X,Y}(\tau) = r_{Y,X}(-\tau)$)

cov. matrix fn.
— Put auto and co's in one matrix

$$\underline{r}_{X,Y}(\tau) = \begin{bmatrix} r_X(\tau) & r_{X,Y}(\tau) \\ r_{Y,X}(\tau) & r_Y(\tau) \end{bmatrix}$$

cross-spectrum
— $R_{X,Y}(f) \triangleq \int_{-\infty}^{\infty} r_{X,Y}(\tau)\, e^{-i2\pi f\tau}\, d\tau$ if $\int |r_{X,Y}(\tau)|\,d\tau < \infty$

matrix
— $\underline{R}_{X,Y}(f) = \begin{bmatrix} R_X(f) & R_{X,Y}(f) \\ R_{Y,X}(f) & R_Y(f) \end{bmatrix}$.

— splitting up $R_{x,y}(f) = A_{x,y}(f)\, e^{i\phi_{x,y}(f)}$

"cross-Amplitude spectrum" ← $A_{x,y}(f) \triangleq \left| R_{x,y}(f) \right| \geq 0$

"phase spectrum" ← $\phi_{x,y}(f) \triangleq \angle R_{x,y}(f) \Rightarrow 0 \leq \phi_{x,y}(f) \leq 2\pi$

"squared coherence spectrum" ← $K^2_{x,y}(f) \triangleq \dfrac{\left| R_{x,y}(f) \right|^2}{R_x(f)\, R_y(f)}$

— Properties:

$$A_{x,y}(-f) = A_{x,y}(f)$$

$$\phi_{x,y}(-f) = -\phi_{x,y}(f)$$

$$0 \leq K^2_{x,y} \leq 1$$

given now without proof

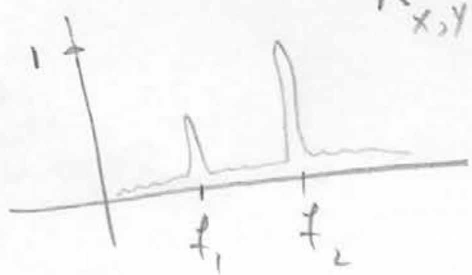
— Interpretations

— $K^2_{x,y}$ — much like correlation function but now showing relations b/w frequency components (instead of time samples)

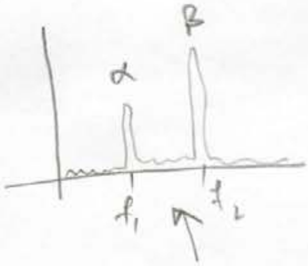
— Plot shows the common/coupled freqs. of X and Y

— Peak heights show how strong is the coupling (proportionality) b/w freqs in X and Y on a scale of 0 to 1 (-ve covered by phase spectrum)

— 1 means maximum linear relation (b/w coupled freqs.) while less than one shows less linear relationship (eg. due to nonlinearity, noisiness etc.)



← shows that only components f_1 and f_2 are coupled b/w X and Y with amplitudes of components at f_2 showing max. linear relationship.



— $A_{x,y}(f)$ ← shows the actual values of the proportionality factors (not just a normalized measure as in $K^2_{x,y}$)

— $\phi_{x,y}(f)$ ← shows the relation between the phases of the coupled/common freq. components of X and Y known from $K^2_{x,y}$ or $A_{x,y}$

f_1 (which is present in both) appears with a phase difference of $\pi/2$

f_2 (" " ") " " " " " " π

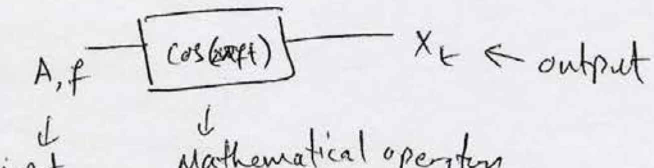
Lec 12 ARMA Models (DT).

(A) What is modeling?

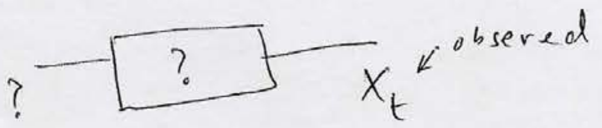
— Efficient way of representing phenomena or information.

— e.g. If we observe X_t as

we can efficiently show that as a model



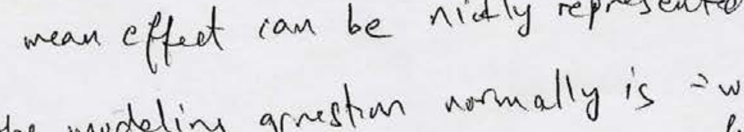
— In terms of stochastic processes



what operation (filter) and input can efficiently represent this observed WSS stochastic process?

— In stochastic processes it is not very important to have exactly the same realization, rather what matters is that what is the mean effect of the process.

— Since that mean effect can be nicely represented by the PSD, the modeling question normally is ⇒ what input and filter can efficiently reproduce this PSD?



(B) Why do modeling?

- they may be required to replicate the behaviour of physical phenomena or systems (e.g human vocal column etc.)
- Represent information efficiently (compression).
- Models allow for interpolation (and in some cases extrapolation).
- Synthesis etc.

(C) How is it done?

(1) The Rational Transfer Function

- In most realistic scenarios we have information of the inputs and output (black box approach)

Input	output
v_0	X_0
v_1	X_1
v_2	X_2
⋮	⋮

- For an LTI system, we may like to write the output as a linear combination of previous inputs and outputs

$c_0 = 1$ conventionally. ; conventionally -ve, can be chosen +ve.

$$X_t = \left(c_0 v_t + c_1 v_{t-1} + \cdots + c_q v_{t-q}\right) - \left(a_1 X_{t-1} + a_2 X_{t-2} + \cdots + a_p X_{t-p}\right) \quad \text{— (A)}$$

First bracket: contains info about input evolution and history. Second bracket: contains info about system evolution (and history)

- e.g if I move a pen on a table, knowing my previous perturbations and pen's previous responses can help me predict how the pencil would behave now and what its new state will be.

— How do we write this in a simple form?

— Meet z^{-1}

— From formulas in your book, w.r.t

$$X_t \Leftrightarrow \mathbb{X}(f)$$
$$X_{t-1} \Leftrightarrow e^{-i2\pi f}\,\mathbb{X}(f)$$

— often for convenience in DT we define $z^{-1} = e^{-i2\pi f}$ and write it in time domain as

$$X_{t-1} = z^{-1} X_t$$

which is actually saying that

$$F\{X_{t-1}\} = z^{-1} F\{X_t\}$$

And we call z^{-1} as the delay operator. e.g.

$$X_{t-p} = z^{-p} X_t$$

— And this makes DT notation a whole lot easier.

— Now we can write (★₁) as

$$X_t = \left(c_0 v_t + c_1 z^{-1} v_t + \dots + c_q z^{-q} v_t\right) - \left(a_1 z^{-1} X_t + \dots + a_p z^{-p} X_t\right)$$
$$= \left(\sum_{k=0}^{q} c_k z^{-k}\right) v_t - \left(\sum_{k=1}^{p} a_k z^{-k}\right) X_t$$

$$\Rightarrow \left(1 + \sum_{k=1}^{p} a_k z^{-k}\right) X_t = \left(\sum_{k=0}^{q} c_k z^{-k}\right) v_t$$

$$\Rightarrow \quad \frac{X_t}{v_t} = \frac{\sum_{k=0}^{q} c_k z^{-k}}{\sum_{k=0}^{p} a_k z^{-k}} \triangleq \frac{C(e^{-i2\pi f})}{A(e^{-i2\pi f})} = H(f)$$

← mixed TD, FD notation. with $c_0 = a_0 = 1$

↓ T.F. of a rational causal LTI system.

(2) PSD of Rational Filter

— As we saw last time for an LTI filter

$$v_t \longrightarrow \boxed{H(f)} \longrightarrow X_t \quad \Rightarrow \quad R_X(f) = |H(f)|^2 R_v(f)$$

(3) Stability of Rational TF

— What's required for $H(f)$ to be stable?

— OF POLES & ZEROS

$$H(f) = \frac{C(e^{-i2\pi f})}{A(e^{-i2\pi f})} \begin{array}{l} \to \text{roots of numerator} = \text{zeros} \\ \to \text{roots of denominator} = \text{poles} \end{array}$$

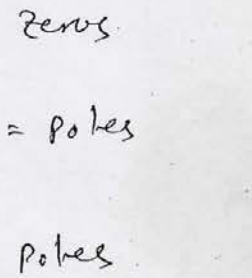
— A DT rational system is stable if all its poles lie inside the unit circle.

— why's that?

or PhD

From theory of DFT (not covered here, known earlier) we know that (considering simple example of single pole)

$$\frac{1}{1 - \alpha z^{-1}} \leftrightarrow \alpha^t \quad \longleftarrow \text{in TD.} \quad \Longrightarrow$$



has a pole at $z = \alpha$

→ stable for $|\alpha| < 1$ which is unit circle for complex α.

(5) ~~(48)~~

(4) White noise PSD As ~~clay~~'. (Input candidate)

— Recall white noise is a process that is completely uncorrelated
— DT white noise e_t $(t = 0, \pm 1, \pm 2, \dots)$ (zero-mean WSS white noise).

$$r_e(\tau) = \begin{cases} \sigma^2 & \tau = 0 \\ 0 & \text{otherwise} \end{cases} \iff R_e(f) = \sigma^2 \qquad f \in (-\tfrac{1}{2}, \tfrac{1}{2}]$$

— White noise has equal contributions from all (allowed) frequencies.

— This has several implications.

(a) White noise is not being 'filtered' through a system (as it is perfectly 'random', so there is no system behind it, or no frequencies are being stopped / magnified).

(b) ~~We can use the white noise PSD as the 'clay' from which to form other WSS PSD's~~ (and white noise processes are easy to generate). (generate an almost random process).

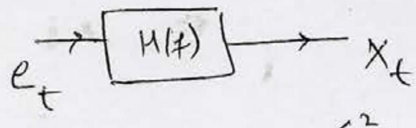
(5) MA

— If you see PSD of white noise signal then there isn't much statistical modeling to do — except giving variance of the white noise, σ^2

— But what if you see this kind of spectrum?

— What happened to the freq. at f_0?
— obviously there is some 'system' that is dropping (zeroing) the freq at f_0.

— let's draw that system



— $R_x(f) = |H(f)|^2 \sigma^2$ ~~$R_e(f)$~~ ($H(f)$ should be zero at $f_0 = 0$)

— A system with only zeros may be written as

$$H(f) = \sum_{k=0}^{q} c_k z^{-k} = C\left(e^{-i2\pi f}\right) \leftarrow \text{has only zeros.}$$

(this is the numerator of the rational TF).

— $R_x(f) = |C(e^{-i2\pi f})|^2 \sigma^2$

— and $X_t = H(f)\, e_t = \sum_{k=0}^{q} c_k e_{t-k}$ $(c_0 = 1)$

— An all-zero system is called $MA(q)$.

— Why the name MA? e.g. for $q = 2$ we might have

$$X_t = e_t + \frac{1}{3} e_{t-1} + \frac{1}{10} e_{t-2}$$

← "kind" of average of last three inputs (moving with (t)) — called "improper averages"

— MA Properties

— Always stable (as they have no poles).

Th. 7.2

— WSS

— PSD: $R_x(f) = \left| \sum_{k=0}^{q} c_k z^{-k} \right|^2 \sigma^2 = r_x(0) + 2\sum_{\tau=1}^{q} r_x(\tau)\cos(2\pi f \tau)$

— Always zero mean (see example)

— covariance fn.

$$r_x(\tau) = \begin{cases} \sigma^2 \sum_{j-k=\tau} c_j c_k & \text{for } |\tau| \le q \\ 0 & ~~\text{otherwise}~~ |\tau| > q \end{cases}$$ ← Noteworthy.

(☆$_2$) —

— e.g. for MA(2)

$$X_t = c_0 e_t + c_1 e_{t-1} + c_2 e_{t-2}$$

$$\Rightarrow m_x = E\{X_t\} = c_0 \underset{\downarrow\,0}{E\{e_t\}} + c_1 \underset{\downarrow\,0}{E\{e_{t-1}\}} + c_2 \underset{\downarrow\,0}{E\{e(t-2)\}}$$

as e_t assumed zero-mean WSS white noise

also $r_x(\tau) = C\{c_0 e_t + c_1 e_{t-1} + c_2 e_{t-2},\ c_0 e_{t+\tau} + c_1 e_{t-1+\tau} + c_2 e_{t-2+\tau}\}$

↓

WSS as e_t is WSS and filter is LTI.

$$\Rightarrow r_x(0) = \sigma^2 (c_0^2 + c_1^2 + c_2^2)$$ ($\swarrow r_e(\tau)$)

$$r_x(1) = \sigma^2 (c_0 c_1 + c_1 c_2) = r_x(-1)$$

$$r_x(2) = \sigma^2 (c_0 c_2) = r_x(-2)$$

$$r_x(\tau) = 0, \quad |\tau| > 2$$

→ May generalize to get (☆$_2$)

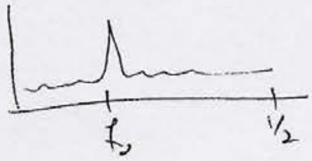
Note: this can also be a criteria for selecting MA as a modeling framew[ork]

↓

— Main feature in time-domain: An MA(q) process has a vanishing covariance. In particular, covariance of an MA(q) process is 0 for $|\tau| > q$.

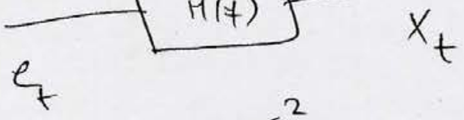
(6) AR

— What if you see PSD like this?



— Some system is magnifying or generating such frequency.

— Let's draw it



— again $R_x(f) = |H(f)|^2 R_e(f)$ ($R_e(f) \to \sigma^2$)

— To magnify things we can use poles (where denominator near gets really small).

$$H(f) = \frac{1}{\sum_{k=0}^{P} a_k z^{-k}} = \frac{1}{A(e^{-i2\pi f})} \quad \left(\text{with } a_0 = 1\right)$$

— This is the deno of rational TF and has poles only.

Th. 7.1 $\left\{ — R_x(f) = \dfrac{\sigma^2}{|A(e^{-i2\pi f})|^2} \right.$

— and $X_t = H(f)\, e_t = -\sum_{k=1}^{P} a_k X_{t-k} + e_t$ ← zero mean WSS white noise as before.

— An all-pole system is called AR.

— why AR?, e.g., for $P=2$ ↓ the name

$$X_t = \overbrace{-a_1 X_{t-1} - a_2 X_{t-2}}^{\text{regressing to its old values}} + e_t$$

↓ i.e. their is a "feedback" of past values into the system.

— AR properties.

— stability requires all poles to be inside unit circle.
— WSS.
— Always zero-mean.
— covariance fn (recursive definition)

Th. 7.1

$$r_x(\tau) + a_1 r_x(\tau-1) + \dots + a_p r_x(\tau-p) = \begin{cases} \sigma^2 & \text{if } \tau = 0 \\ 0 & \text{if } \tau \neq 0 \end{cases}$$

— these are called Yule-Walker equations (and need to be solved to get $r_x(\tau)$).

— Proof of zero-mean → WSS

$$X_t = -\sum_{k=1}^{p} a_k X_{t-k} + e_t$$

$$\Rightarrow \underset{m_x}{E\{X_t\}} = -\sum_{k=1}^{p} a_k \underset{m_x}{E\{X_{t-k}\}} + E\{\cancelto{0}{e_t}\}$$

$$m_x + a_1 m_x + \dots + a_p m_x = 0$$

$$m_x \underbrace{(1 + a_1 + \dots + a_p)}_{\neq 0} = 0$$

b/c if this is zero then the ~~polynomial~~ TF we are dealing with, i.e., $1/\left(\sum_{k=0}^{p} a_k z^{-k}\right)$ has a pole at $z=1$ which is not allowed by stability restriction (poles should be inside unit circle).

— see book ~~pg 134~~ for Yule-Walker.

- why use AR?

  - Many series (processes) are actually generated via feedback systems.

  - AR-Model is very flexible and can cover a large range of covariance & spectrum structures.

  - Parameter estimation is simple (coming up next)

  - Good choice for forecasting (e.g.)

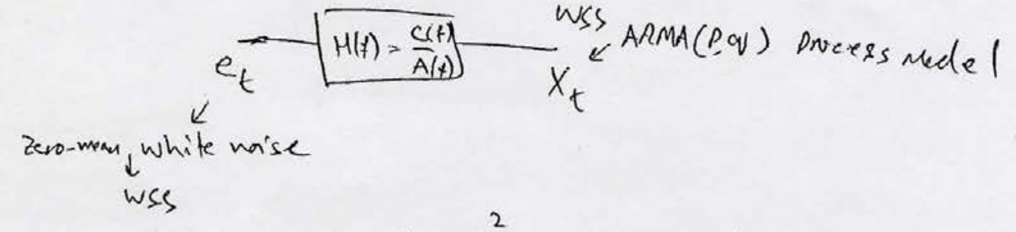
known past

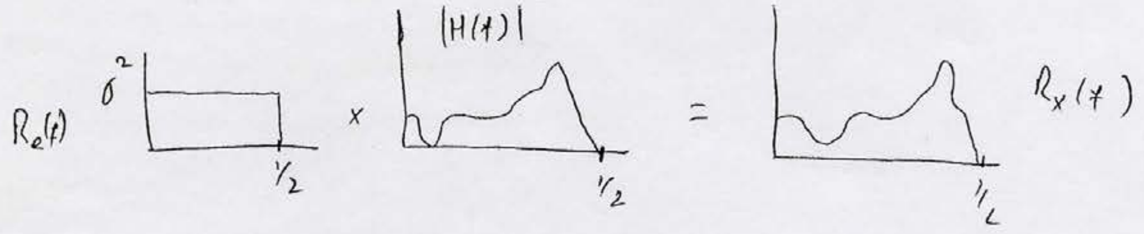
$$\hat{X}_{t+1} = -a_1 X_t - a_2 X_{t-1} - \cdots - a_p X_{t-p+1}$$

future prediction — estimated params

(7) Link b/w Pole-Zero Plots and PSD.

— We saw last time that we can model stationary st. processes using the rational TF with white noise input.





— These are related mathematically as

$$R_x(f) = |H(f)|^2 R_e(f)$$

$$\Rightarrow \quad R_x(f) = \frac{|C(e^{-i2\pi f})|^2}{|A(e^{-i2\pi f})|^2}\sigma^2$$

$$= \frac{\left|\sum_{k=0}^{q} c_k z^{-k}\right|^2}{\left|\sum_{k=0}^{p} a_k z^{-k}\right|^2}\sigma^2 \quad \text{with } c_0 = a_0 = 1$$

— Peaks with Poles and Nulls with Zeros

$$H(f) = \frac{\sum_{k=0}^{q} c_k z^{-k}}{\sum_{k=0}^{p} a_k z^{-k}}$$

Numerator $\rightarrow$ a polynomial in z whose roots form zeros of $H(f)$

Denominator $\rightarrow$ " " " " " " Poles of $H(f)$

— How do you tell the link b/w pole-zero locations and PSD shapes?
(show figures).

— we consider an MA(2) example (results generalize analogously)

$$H(f) = \sum_{k=0}^{2} c_k z^{-k} = 1 + c_1 z^{-1} + c_2 z^{-2} = C(z^{-1})$$

— for convenience we can rewrite it as a polynomial in z.

$$C(z^{-1}) = 1 + c_1 z^{-1} + c_2 z^{-2}$$

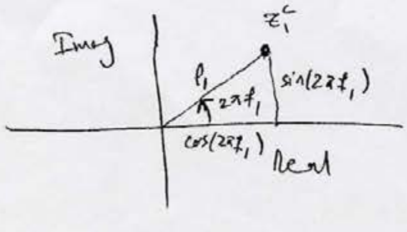
set to zero to find roots.

$$\Rightarrow z^2 C(z^{-1}) = z^2 + c_1 z + c_2 = 0 \quad \leftarrow \text{called ch. eqn.}$$

Now suppose that we find roots of ch. eqn to be

z_1^c, z_2^c → can be complex in general

so $z_1^c = \rho_1 e^{i2\pi f_1}$, $z_2^c = \rho_2 e^{i2\pi f_2}$ ← Polar notation of complex numbers



$$z_1^c = \rho_1 \cos(2\pi f_1) + i\rho_1 \sin(2\pi f_1) = \rho_1 e^{i2\pi f_1}$$

(note, for unit circle set $\rho_1 = 1$)

— ch. eqn can now be written as

$$z^2 c(z^{-1}) = (z - z_1^c)(z - z_2^c)$$

$$c(z^{-1}) = \frac{(z - z_1^c)(z - z_2^c)}{z^2}$$

— now the PSD of this MA(2) process is

$$R_x(f) = \left|\sum_{k=0}^{2} c_k z^{-k}\right|^2 \sigma^2$$

$$= \frac{|z - z_1^c|^2 \, |z - z_2^c|^2}{|z^2|^2} \sigma^2$$

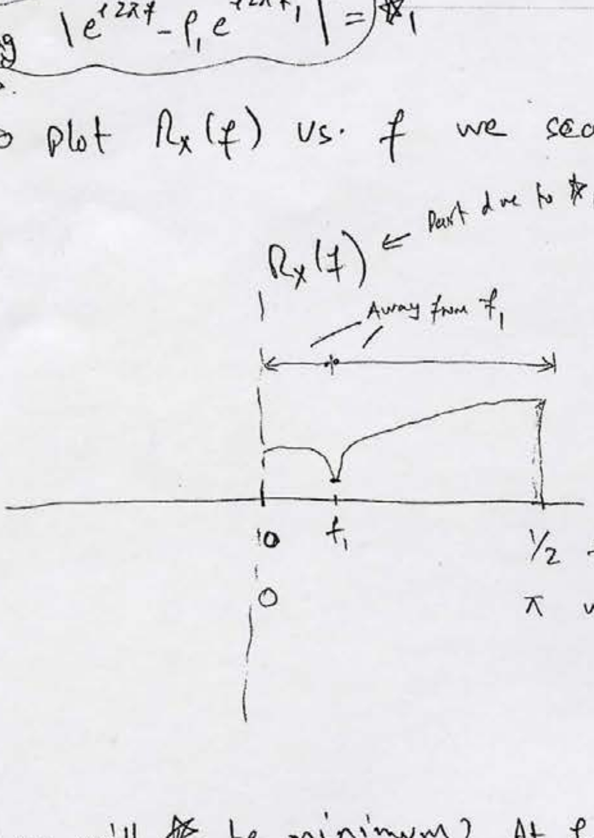
since $z = e^{i2\pi f}$ (as used in FT), and $z_1^c = \rho_1 e^{i2\pi f_1}$, $z_2^c = \rho_2 e^{i2\pi f_2}$

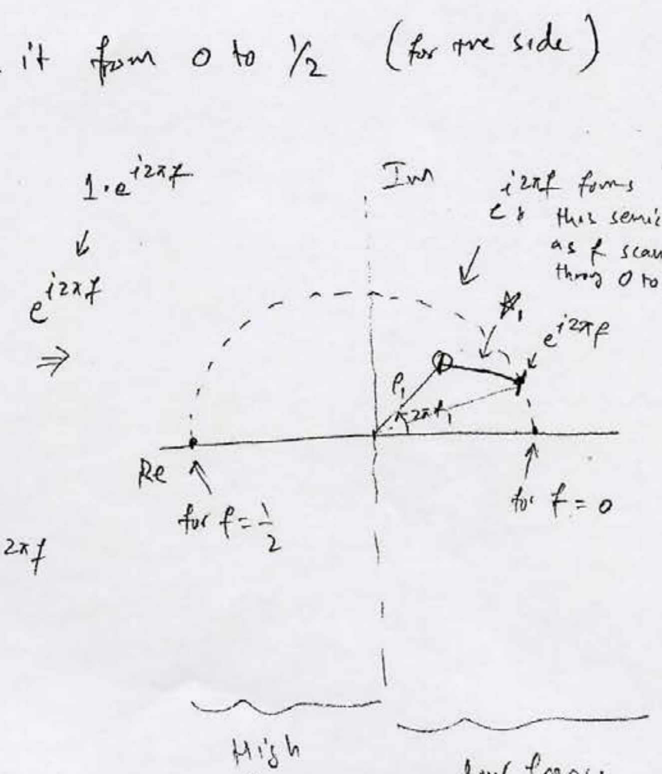
$$\Rightarrow \quad R_x(f) = \frac{|e^{i2\pi f} - \rho_1 e^{i2\pi f_1}|^2 \, |e^{i2\pi f} - \rho_2 e^{i2\pi f_2}|^2}{|e^{i2\pi f}|^2} \sigma^2$$

($|e^{i2\pi f}|^2 \to 1$)

(studying) $|e^{i2\pi f} - \rho_1 e^{i2\pi f_1}|^2 = ☆_1$

— To plot $R_x(f)$ vs. f we scan it from 0 to ½ (for +ve side)





— where will ☆₁ be minimum? At f_1.

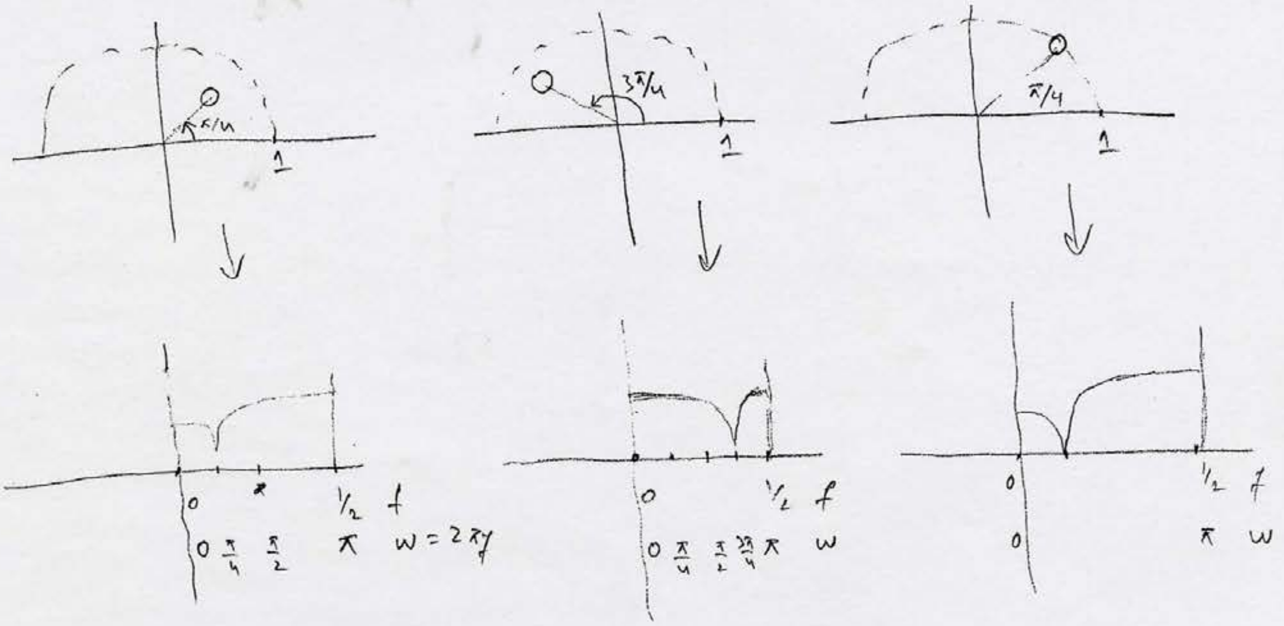
— what is the value of this minimum?

$$☆_1 = |e^{i2\pi f_1} - \rho_1 e^{i2\pi f_1}|^2 = |1 - \rho_1|^2 \, |e^{i2\pi f_1}|$$

$$= |1 - \rho_1|^2 \Rightarrow$$ The minimum of ☆₁ is deeper when ρ_1 is closer to 1 (i.e. when z_1^c lies closer to zero on unit circle)

$2\pi/4 + \frac{\pi}{4} = \frac{3\pi}{4}$

— How can you connect the plots?



— Similar holds for AR (and more generally ARMA) models.

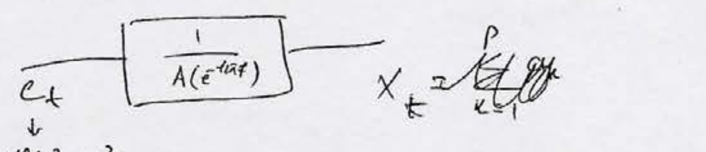
— remember that stability requires poles to be inside unit circle.

— ~~often~~ It's also wiser to have zeros inside unit circle
(b/c we may want to invert the model, zeros outside unit circle would become poles outside unit circle) $\rightarrow$ $z - z_1 \rightarrow \frac{1}{z - z_1}$

(8) AR coefficient estimation

— We talk of AR only. General ARMA are a bit more tricky (covered in Time Series course that follows this one).

— In cases where the covariance function is known or is convenient to estimate from data, we make use of YWE.



— We need to estimate the a_k coefficients and σ^2 (then any white noise with zero-mean & σ^2 would do)

— By YWE (assuming $r_x(\tau)$ known or already estimated from data)

$$r_x(\tau) + a_1 r_x(\tau-1) + \dots + a_p r_x(\tau-p) = \begin{cases} \sigma^2 & \tau = 0 \\ 0 & \tau \neq 0 \end{cases}$$

— In matrix form (for $\tau \neq 0$)

using $r_x(-1) = r_x(1)$ etc.

$$\underbrace{\begin{bmatrix} r_x(0) & r_x(1) & \dots & r_x(p-1) \\ r_x(1) & r_x(0) & \dots & r_x(p-2) \\ \vdots & \vdots & & \vdots \\ r_x(p-1) & r_x(p-2) & \dots & r_x(0) \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}}_{\underline{a}} = - \underbrace{\begin{bmatrix} r_x(1) \\ r_x(2) \\ \vdots \\ r_x(p) \end{bmatrix}}_{\underline{r}_1}$$

($AX = b$ form; A known, b known)

$$\Rightarrow \Sigma \underline{a} = -\underline{r}_1 \Rightarrow \hat{\underline{a}}^* = -\Sigma^{-1} \underline{r}_1$$

↑ Provides estimate of the a_k's.

— Estimate σ^2 using $\hat{\underline{a}}^*$ and YWE for $\tau = 0$

$$\Rightarrow \hat{\sigma}^{*2} = r_x(0) + \hat{a}_1^* r_x(1) + \dots + \hat{a}_p^* r_x(p)$$

— This is called the autocorrelation method.

— sometimes its more convenient to estimate the coefficients directly from the data.

— consider AR(p)

$$X_t + a_1 X_{t-1} + \cdots + a_p X_{t-p} = e_t$$

— we have available observations $x_1, x_2, x_3, \ldots, x_n$

— each one should fit the AR(p) model.

$$x_1 + a_1 x_0 + \cdots + a_p x_{1-p} = e_1$$ (unavailable)

$$x_2 + a_1 x_1 + \cdots + a_p x_{2-p} = e_2$$

$$\vdots$$

$$x_p + a_1 x_{p-1} + \cdots + a_p x_0 = e_p$$ (unavailable)

$$\star \begin{cases} x_{p+1} + a_1 x_p + \cdots + a_p x_1 = e_{p+1} \\ x_{p+2} + a_1 x_{p+1} + \cdots + a_p x_2 = e_{p+2} \\ \vdots \\ x_n + a_1 x_{n-1} + \cdots + a_p x_{n-p} = e_n \end{cases}$$ ← all x's available.

→ In Matrix form, ⊛ can be written as

$$\underbrace{\begin{bmatrix} x_{p+1} \\ x_{p+2} \\ \vdots \\ x_n \end{bmatrix}}_{\underline{X}} = \underbrace{\begin{bmatrix} -x_p & -x_{p-1} & \cdots & -x_1 \\ -x_{p+1} & -x_p & \cdots & -x_2 \\ \vdots & \vdots & & \vdots \\ -x_{n-1} & -x_{n-2} & \cdots & -x_{n-p} \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix}}_{\underline{a}} + \underbrace{\begin{bmatrix} e_{p+1} \\ e_{p+2} \\ \vdots \\ e_n \end{bmatrix}}_{\underline{E}}$$

generally $n >> p$ (overdetermined system i.e. more equations than unknowns)

$$\Rightarrow \underline{X} = U\underline{a} - \underline{E}$$

→ find $\underline{a}^*$ that minimizes $|E|^2$ ← minimize the white noise dynamics that do not fit a linear regression ← so $\underline{a}$ is chosen to fit a linear regression as much as possible.

$$Q(\underline{a}) = |\underline{X} - U\underline{a}|^2 = (\underline{X} - U\underline{a})'(\underline{X} - U\underline{a})$$

→ $\underline{a}^* = \arg\min_{\underline{a}} Q(\underline{a})$

→ Taking $\frac{\partial Q(\underline{a})}{\partial \underline{a}}$ and using matrix algebra (skipped!), we get

$$\underline{a}^* = (U'U)^{-1} U'\underline{X}$$ ← called the LS solution. (since $Q(\underline{a})$ is in squares form)

→ covered in more detail in Time series analysis course.

— $\hat{\sigma}^2 = \frac{Q(\underline{a}^*)}{n-p}$ (since $Q(\underline{a}) = |E|^2 = E'E = e_{p+1}^2 + e_{p+2}^2 + \dots + e_n^2 = \sum_{k=p+1}^{n} e_k^2$) ← recall that e_k is zero-m.

— Due to this, the method is called "covariance method" ← note

(9) Prediction.

— once a_k's are known, we can estimate future values using the AR(p) model

$$X_{t+1} = -a_1^* X_t - a_2^* X_{t-1} + \dots - a_p X_{t-p+1} + e_{t+1}$$

(a's estimated; $X_t, X_{t-1}, \dots$ observed; X_{t+1} future; e_{t+1}:)
→ uncorrelated with know $X_t, X_{t-1} \dots$
→ cannot be estimated from know $X_t, X_{t-1} \dots$
→ best we can do is to replace it by its expected value i.e. $E\{e_{t+1}\} = 0$

$$\Rightarrow X_{t+1}^* = -a_1^* X_t - a_2^* X_{t-1} - \dots - a_p^* X_{t-p+1}$$

Lec ~~to~~ (13) (1)

(1) CT Linear Filters

(1.1) Linear stochastic Differential equations (analogous to AR(p))

— we have seen the "difference" equation for AR(p)

$$a_0 Y_t + a_1 Y_{t-1} + \dots + a_p Y_{t-p} = X_t \quad \swarrow \text{Input}$$

— And we used $z^{-1} = e^{-i2\pi f}$ to write

$$Y_{t-1} = z^{-1} Y_t \iff F\{Y_{t-1}\} = e^{-i2\pi f} F\{Y_t\}$$

— And we found it TF to be

$$Y_t = \overset{H(f)}{\boxed{\frac{1}{\sum_{k=0}^{p} a_k z^{-k}}}} X_t \iff Y(f) = H(f) X(f)$$

— And output PSD as

$$R_y(f) = |H(f)|^2 R_x(f)$$

— Now we see in CT, where "difference" is replaced by "differential"

$$a_0 Y^{(p)}(t) + a_1 Y^{(p-1)}(t) + \dots + a_{p-1} Y'(t) + a_p Y(t) = X(t)$$

where $Y^{(p)}(t) = \frac{d^p}{dt^p} Y(t)$.

— Introduce $r = i2\pi f$

Known property of differential FT.

$$Y'(t) = r Y(t) \iff F\{Y'(t)\} = (i2\pi f) F\{Y(t)\}$$

— we have

$$\left(a_0 r^p + a_1 r^{p-1} + \dots + a_{p-1} r + a_p\right) y(t) = x(t)$$

$$\Rightarrow \quad y(t) = \boxed{\frac{1}{\sum_{k=0}^{p} a_k r^{p-k}}} \, x(t) \qquad \Longleftrightarrow \qquad Y(f) = H(f)\, X(f)$$

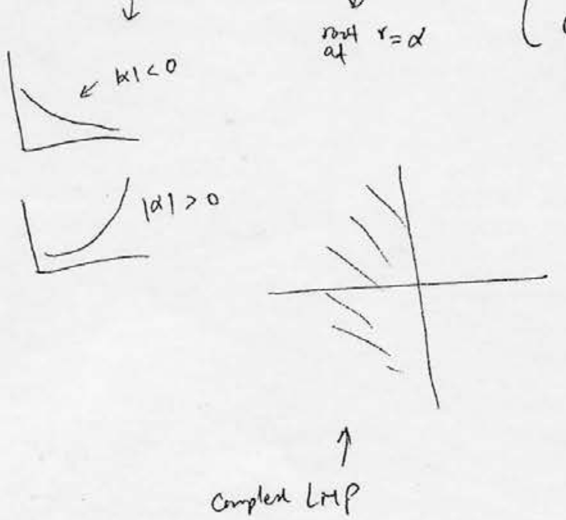
↘ $H(f)$ (with $r = i2\pi f$)

— $H(f)$ stable if all roots are in LHP.

— b/c $\quad e^{+\alpha t} \overset{F.T}{\Longleftrightarrow} \dfrac{1}{r-\alpha} \quad$ (for $r = i2\pi f$) ← C.T

↓ root at $r = \alpha$

(Recall $\alpha^n \longleftrightarrow \dfrac{1}{z-\alpha}$) ← for D.T. $z = e^{i2\pi f}$



— For stable system

$$R_y(f) = |H(f)|^2 R_x(f) = \frac{1}{\left|a_p + a_{p-1}(i2\pi f) + \dots + a_0 (i2\pi f)^p\right|^2} R_x(f)$$

$$Y(f) = H(f)\, X(f)$$

by causality of our definition. ($h(\tau) = 0 \ \forall\ \tau < 0$)

$$\Rightarrow \quad y(t) = \int_{-\infty}^{\infty} h(t-u)\, x(u)\, du = \int_{-\infty}^{t} h(t-u)\, x(u)\, du$$

— further it can be shown the The IR $h(t)$ satisfies

$$a_0 h^{(p)}(t) + a_1 h^{(p-1)}(t) + \dots + a_{p-1} h'(t) + a_p h(t) = 0 \quad \text{for } t \neq 0$$

with IC's

$$h(0) = h'(0) = \dots = h^{(p-2)}(0) = 0$$

and $h^{(p-1)}(0) = \frac{1}{a_p}$.

Recall that soln of a pth order ODE requires $(p-1)$ initial conditions.

4

(2) An important CT filter : Hilbert Tx.

— Recall that we use complex functions b/c they simplify analysis. e.g. prove $\cos x \cdot \cos y = \frac{1}{2}\left(\cos(x+y) + \cos(x-y)\right)$
→ easy if we use $\cos x = \frac{e^{ix} + e^{-ix}}{2}$.

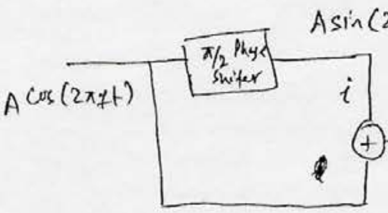
— As I mentioned before, real-world processes/signals are real.

— Then where do the complex functions come from?

— e.g. for cosine signal we get the complex one like this

makes manipulation easier ↓

analytic signal ↓

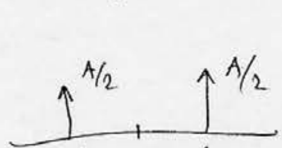


$\rightarrow A(\cos 2\pi f t + i \sin 2\pi f t) = A e^{i 2\pi f t}$

(note: replacing $2\pi f t$ with $2\pi f t + \pi/2$ is NOT same as replacing t by $t + \pi/2$
↓
i.e. a simple time-translation of signal doesn't work).

— notice that the real signal has both +ve & −ve freqs

$A\cos(2\pi f t) \Rightarrow$



← FT

— while the complex one has only +ve freq

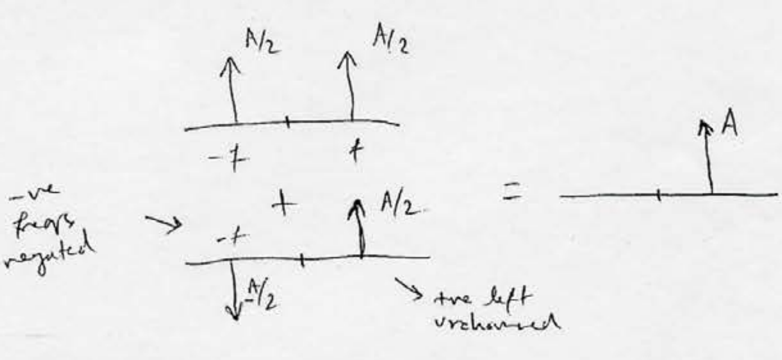
$A e^{i 2\pi f t} \Rightarrow$

← thus we are dealing only with +ve frequencies. (useful in many applications).

— And the two have same Power

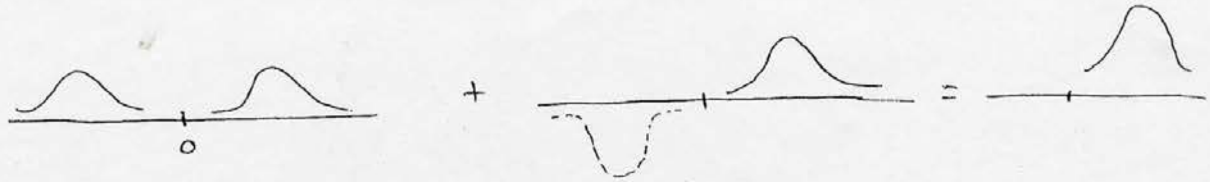
— And we can recover the original from analytic.

— So, what did we really do to get the nice analytic signal?



— now what abt a stochastic signal $X(t)$ that may have several freqs?
how do we get its analytic signal version (say $X^+(t)$ ← complex counterpart of real $X(t)$)

— What we want is (analogous to simple cos case)



— we want $X^+(t) = X(t) + i\,Y(t)$ where

$$\left.\begin{aligned} X^+(f) &= 0 = X(f) - X(f) & f<0 \\ X^+(f) &= 2X(f) = X(f) + X(f) & f>0 \\ X^+(f) &= X(f) & f=0 \end{aligned}\right\} \circledast$$

— but $X^+(f) = X(f) + i\,Y(f)$ (just take FT of $X^+(t)$)
— comparing ↑ with ⊛, case by case, gives

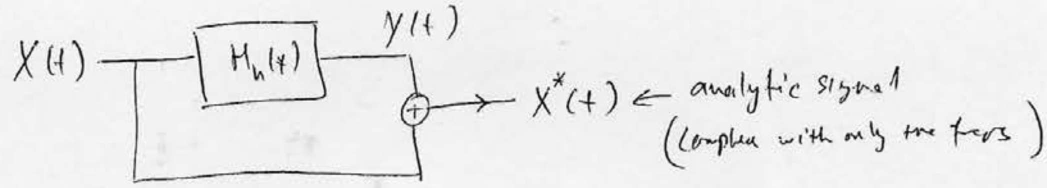
$$Y(f) = \begin{cases} i\,X(f) & f<0 \\ -i\,X(f) & f>0 \\ 0 & f=0 \end{cases}$$

— writing this in filter form

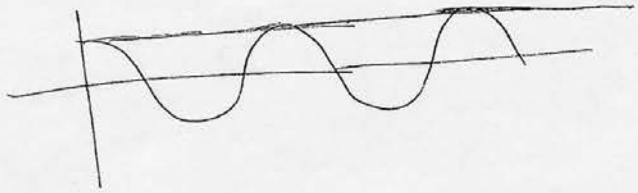
$$H_h(f) = \begin{cases} i & f<0 \\ -i & f>0 \\ 0 & f=0 \end{cases}$$

$$Y(f) = H_h(f)\,X(f)$$

— This is the Hilbert Tx, which provides the analytic signal for a given stationary real signal (very useful)

— 

— just like $|Ae^{i2\pi ft}|$ gives the "envelope" of $A\cos 2\pi ft$ (which is $|A|$)

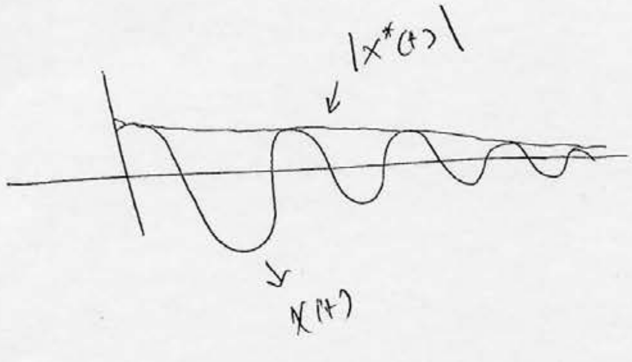


$Z(t) \triangleq |X^*(t)|$ gives the envelope of $X(t)$.

since ~~for~~ $X^*(t) = X(t) + i\,y(t) = $ ~~$\sqrt{X^2(t)+y^2(t)}\,e$~~ $|X^*(t)|\,e^{i\angle X^*(t)}$

↓ magn.

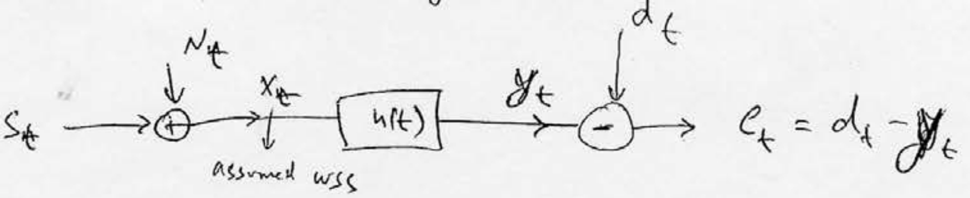
eg



— Hilbert Tx can thus be used as an envelope extractor as well.

— optimal filters

— we talked about voice corrupted by noise



— we want max S_t and min N_t at output

— This is called "optimization", and a filter doing this is called an "optimal" filter.

— we consider a more general case with a general desired output d_t

— The output "error" is

$$e_t = d_t - y_t = d_t - \sum_{l=-\infty}^{\infty} h_l x_{t-l}$$

(notice that this can be non-causal (causal in a future course!))

— One way to choose the filter, $h(t)$, is to choose it such that the mean squared error is minimized

$$\{h_t\} = \arg\min_{\{h_t\}} E\{e_t^2\}$$

— ∴ take derivative and set to zero

$$\frac{\partial}{\partial h_t} E\{e_t^2\} = E\left\{2 e_t \frac{\partial}{\partial h_t} e_t\right\}$$

$$= 2E\left\{ e_t \left[\underbrace{\frac{\partial}{\partial h_t} d_t}_{=0} - \sum_\ell \underbrace{\left(\frac{\partial}{\partial h_t} h_\ell\right)}_{\substack{=0 \text{ if } k\neq \ell \\ =1 \text{ if } k=\ell}} x_{t-\ell} \right] \right\}$$

$$= -2E\left\{ e_t x_{t-k} \right\} = 0$$

$$\Rightarrow \quad E\left\{ e_t x_{t-k} \right\} = 0$$

$$\Rightarrow \quad E\left\{ \left(d_t - \sum_\ell h_\ell x_{t-\ell}\right) x_{t-k} \right\} = 0$$

$$\Rightarrow \quad \underbrace{E\left\{ d_t x_{t-k} \right\}}_{r_{dx}(k)} - \sum_\ell h_\ell \underbrace{E\left\{ x_{t-\ell} x_{t-k} \right\}}_{r_x(\ell - k)} = 0$$

$$\Rightarrow \quad r_{dx}(k) = \sum_\ell h_\ell r_x(\ell - k) \quad , \; -\infty < k < \infty$$

— still need to solve for $\{h_\ell\}$

— Easier with FT.

$$R_{dx}(f) = H(f)\, R_x(f)$$

$$\Rightarrow \quad H(f) = \frac{R_{dx}(f)}{R_x(f)}$$ ← this is called the "Wiener" filter.

— For the special (more typical) case

$$d_t = S_t \quad \text{and} \quad N_t \perp S_t$$

— Then,

$$r_{dx}(\kappa) = E\{d_t\, x_{t-\kappa}\} = E\{S_t(S_{t-\kappa} + N_{t-\kappa})\}$$

$$= E\{S_t S_{t-\kappa}\} + \underbrace{E\{S_t N_{t-\kappa}\}}_{=0} = r_s(\kappa)$$

— and,

$$r_x(\kappa) = E\{x_t\, x_{t-\kappa}\} = E\{(S_t + N_t)(S_{t-\kappa} + N_{t-\kappa})\}$$

$$= E\{S_t S_{t-\kappa}\} + E\{N_t N_{t-\kappa}\} = r_s(\kappa) + r_N(\kappa)$$

— From above two,

$$R_{dx}(f) = R_s(f)$$

and $R_x(f) = R_s(f) + R_N(f)$

$$\Rightarrow H(f) = \frac{R_s(f)}{R_s(f) + R_N(f)}$$

— Further, it can be shown that the SNR achieved in this case is

$$SNR_{max} = \frac{\int R_s(f)\, df}{\int \frac{R_s(f)\, R_N(f)}{R_s(f) + R_N(f)}\, df}$$

$$\left(SNR = \frac{\text{signal power}}{\text{noise power}} \right)$$

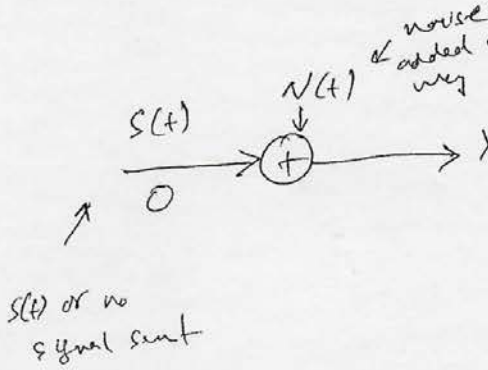
Part B

(10) (69)

- The wiener filter is used to extract a signal out of noise when signal stats are known.

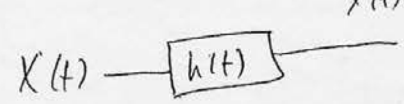
- Another type of requirement can be ~~that~~ "detection" where the question is "does this particular signal/signature appear in the signal"? (NQR, radar etc.)

- consider a communication scenario where either a 0 or 1 is sent.



$$X(t) = \begin{cases} S(t) + N(t) \\ N(t) \end{cases}$$ ← reception scenarios

- suppose we want to apply a filter to the ~~Rx~~ recvd signal



$$Y(t) = \begin{cases} S_{out}(t) + N_{out}(t) \\ N_{out}(t) \end{cases}$$

- The "Detection" or "Decision" problem is → H_0 : "zero" sent, H_1 : "one" sent

mathematically (T = decision instant or time)

$$H_0 : Y(T) = N_{out}(T)$$
$$H_1 : Y(T) = S_{out}(T) + N_{out}(T)$$

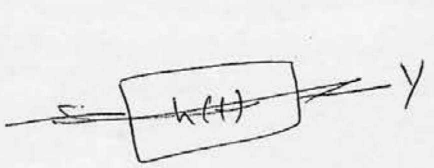
— All we have available for decision is $Y(T)$

— Can we deduce something from the stat of $Y(T)$ for the two possibilities?

— We will have to make some practical assumptions first.

— Let's assume N_t is Gaussian with zero mean and cov. fn. $r_N(\tau)$, and that we choose an LTI filter $h(t)$, and $s(t)$ is deterministic. (e.g. a pulse)

~~— Then $N_{out}(T)$~~



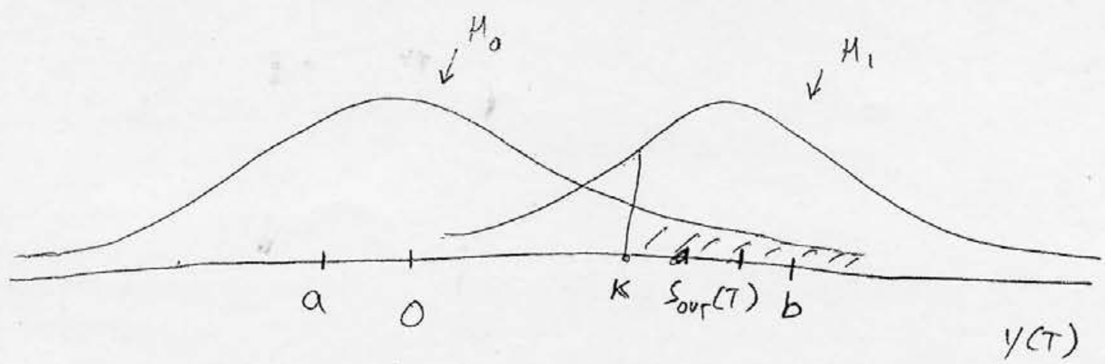
by Gaussian-in-Gaussian-out and $\sigma_N^2(T) = V\{N_{out}(T)\}$
$$= \int_0^T\int_0^T h(u)h(v)\, r_N(u-v)\, du\, dv.$$

— Then ~~$Y(T)$~~

$$H_0 : Y(T) = N_{out}(T) \in \mathcal{N}(0, \sigma_N^2)$$
$$H_1 : Y(T) = S_{out}(T) + N_{out}(T) \in \mathcal{N}(S_{out}(T), \sigma_N^2)$$

↓ we assume $S_{out}(T) > 0$

— Let's plot the distributions



— Suppose $y(T) = a$, what would be a better decision probabilistically?

— Suppose $y(T) = b$, " " " " " ?

— Suppose we make the following decision criterion

Test	Decision
$y(T) \le K$	"Zero" sent
$y(T) > K$	"one" sent

— Is this a fully guaranteed-to-be-correct criterion?

— What is the probability that $y(T) > K$ under H_0?

Let's call this α

— And what is the probability that $y(T) < K$ under H_1?

Let's call this β

wrongly thinking that the signal is there.

— Clearly $\alpha = P\left\{\text{Decide } H_1 \text{ when } H_0 \text{ is true}\right\} = P\left\{\text{False alarm}\right\}$

$$\beta = P\left\{\text{Decide } H_0 \text{ when } H_1 \text{ is true}\right\} = P\left\{\text{Miss}\right\}$$

— And of course we want both α and β to be small.

— How can the filter help us in this decision making?

— would it help us if $S_{out}(T)$ were large? (Yes, as H_1 would move farther to right)

— would it help us if $\sigma_N^2(T)$ were small? (Yes, as the tails would decay faster, giving less power in overlapping distributions)

— then why not make a filter that does both these?

— SNR covers both these, as $SNR = \dfrac{S_{out}^2(T)}{\sigma_N^2(T)}$

it increases by increasing $S_{out}^2(T)$ and decreasing $\sigma_N^2(T)$.

— so we choose a filter that maximizes SNR.

i.e.
$$h_t = \arg\max_{h_t} \frac{\overset{S_{out}^2(T)}{\left[\int s(T-u)\,h(u)\,du\right]^2}}{\underset{\sigma_N^2(T)}{\iint h(u)\,h(v)\,r_N(u-v)\,du\,dv}}$$

— Such a filter is called a "matched" filter (tries to make "matching" easier).

— writing α, β mathematically

$$\alpha = P\left\{ Y(T) > K \mid Y(T) \in \mathcal{N}(0, \sigma_N^2) \right\} = 1 - \Phi\left(\frac{K}{\sigma_N(T)}\right)$$

$$\beta = P\left\{ Y(T) < K \mid Y(T) \in \mathcal{N}(S_{out}(T), \sigma_N^2) \right\} = \Phi\left(\frac{K - S_{out}(T)}{\sigma_N(T)}\right)$$

$$= 1 - \Phi\left(\frac{S_{out}(T) - K}{\sigma_N(T)}\right)$$

— Selection of K affects α and β and scenario dictates which one is more important

— e.g. for landmine detection you would want β to be small.
for airport narcotics detection you would want α to be small.

— For the case when we want $\alpha = \beta$ we can set

$$K = S_{out}(T)/2$$

$$\Rightarrow \alpha = \beta = 1 - \phi\left(\frac{S_{out}(T)}{2\sigma_N(T)}\right)$$

— notice again that both α & β would decrease if $S_{out}(T)$ increase and $\sigma_N(T)$ decreases. (← hence the SNR maximization criteria of matched filter).

— Difference from wiener is here S is deterministic and we use its shape (not statistics) to solve for h_t.

— Need : A large variety of methods for analyzing, modeling & filtering stochastic processes are based on the PSD of the process measured in noise

— All of the above would benefit from a good estimate of the PSD.

(1) Life is short (and so is the data).

— so far we've discussed the theoretical way of finding the PSD, which is to take the FT of the cfn.

$$R_x(f) = \sum_{\tau=-\infty}^{\infty} r_x(\tau) e^{-i2\pi f \tau} \qquad \text{(DT case)}$$

— what's wrong here? — $r_x(\tau)$ is required for all τ, which means we need infinite amount of data to find $R_x(f)$!

— what if we only have data for $0, \dots, n-1$?

Mr. A Gets the "sample" estimate of $r_x(\tau)$ as

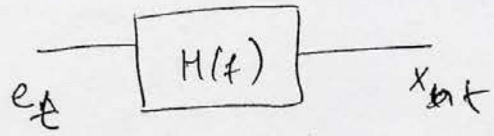
$$r_x^*(\tau) = \frac{1}{n} \sum_{t=0}^{n-1-|\tau|} x(t)\, x(t+|\tau|) \qquad \left(\text{from (2.17) with } m=0\right)$$

and then takes its FT (range allowed by available data)

$$R_x^*(f) = \sum_{\tau=-n+1}^{n-1} r_x^*(\tau)\, e^{-i2\pi f \tau}$$

→ This is the "Periodogram" approach.

Mr.B Uses techniques of chapter 6 to first get an ARMA model estimate of the available process data



$$\hat{H}^*(f) = \frac{\hat{C}^*(e^{-i2\pi f})}{\hat{A}^*(e^{-i2\pi f})}$$ ← estimate polynomials A & B to fit the observed data

Then $R_x^*(f) = |H^*(f)|^2 \cdot \sigma^2$ 1

→ This is a "parametric" approach (as here we first find parameters of a model)

→ Periodogram is a "non-parametric" approach.

→ We discuss here only the periodogram approach and it's variants. (There are also other approaches)

(2) The Periodogram

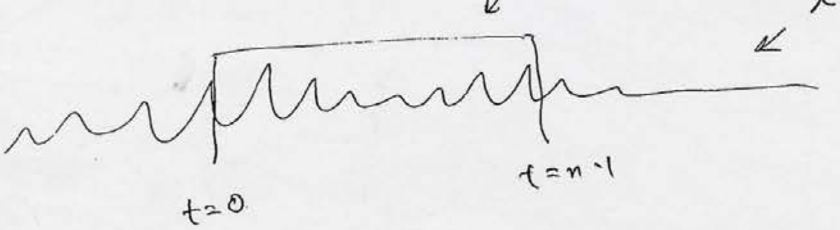
— As noted above, the periodogram is given as

$$R_x^*(f) = \sum_{\tau=-n+1}^{n-1} r_x^*(\tau) e^{-i2\pi f \tau}$$

— Another way of reaching here is the following

— We can see the available data as a 'windowed' segment of the infinite data

$w(t) = \begin{cases} 1 & 0 \le t < n \\ 0 & \text{otherwise} \end{cases}$ ← rectangular window.



— call this segment $x_w(t)$

$$\Rightarrow \quad x_w(t) = x(t)\, w(t)$$

Take cvfn. $\Rightarrow$

$$r_x^*(\tau) = \frac{1}{n} \sum_{t=-\infty}^{\infty} x_w(t)\, x_w(t+\tau)$$

$$= \frac{1}{n} \Big(x_w(t) * x_w(-t) \Big)$$

← by defn. of convolution $f(t) * g(t) = \sum f(t) g(t-\tau)$

Take FT to get PSD $\Rightarrow$

$$R_x^*(f) = \frac{1}{n} X_w(f)\, \overline{X}_w(f)$$

$$= \frac{1}{n} \left| X_w(f) \right|^2$$

$$= \frac{1}{n} \left| \sum_{t=-\infty}^{\infty} x(t)\, w(t)\, e^{-i2\pi f t} \right|^2$$

i.e. → The periodogram is formed from a (windowed FT) of the data.

(3)

Performance Measures

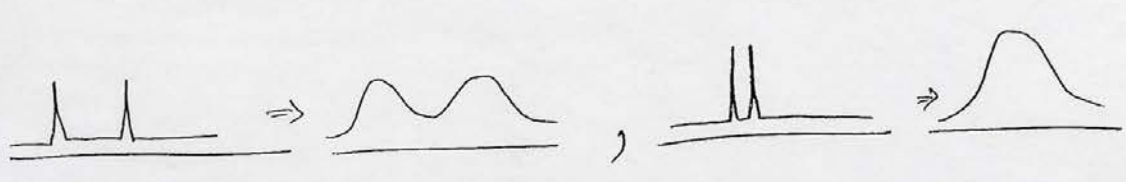
— How do we compare different PSD estimators?

(i) bias $E\{R_x^*(f)\} \stackrel{?}{=} R_x(f)$

(ii) Asymptotically unbiased $\lim_{n \to \infty} E\{R_x^*(f)\} \stackrel{?}{=} R_x(f)$

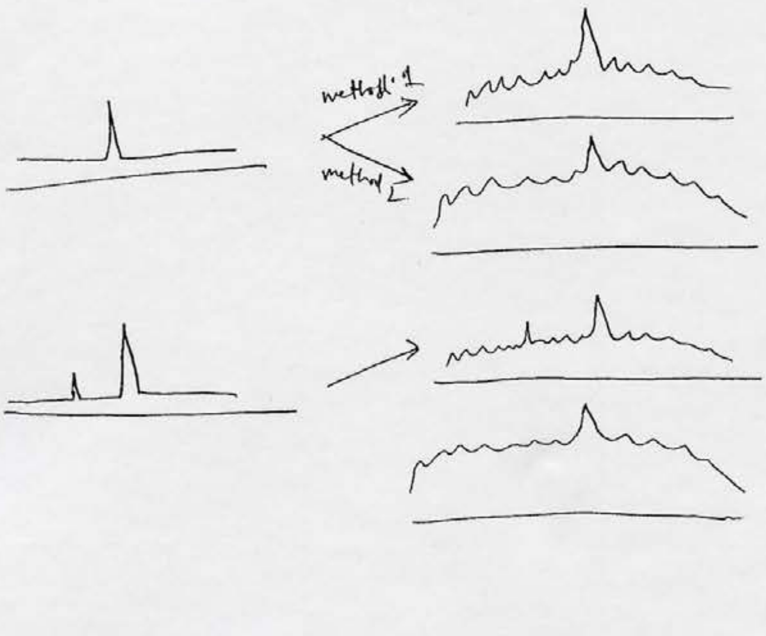
(iii) Consistency, $\lim_{N \to \infty} V\{R_x^*(f)\} \stackrel{?}{=} 0$

(iv) Resolution: ability to show closely spaced peaks separately (sharpness)



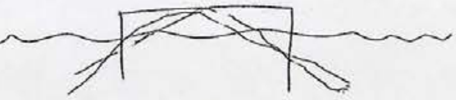
((low res example) (how sharpness relates to resolution)

(v) Masking/leakage: Existence of artifacts that mask (hide) the low power peaks.



← low peak masked by high leakage.

(4) Great Expectations (the ideal window)



— notice that we could window the data in many ways

— It can be shown that for a general window (say $g(t)$) we have

$$\hat{R}_x(f) = \frac{1}{n} \left| \sum_{t=-\infty}^{\infty} x(t) g(t) e^{-i 2\pi f t} \right|^2$$

and

$$E\{\hat{R}_x^*(f)\} = \int_{-1/2}^{1/2} R_x(u)\, G(f-u)\, du$$

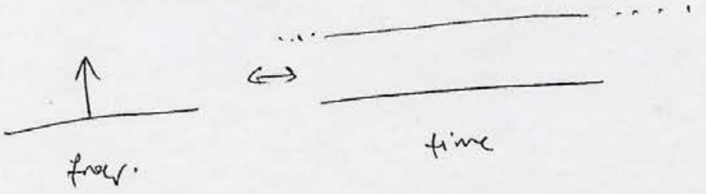
($\pm 1/2$ ← for DT case; G ← FT of $g(t)$)

— Based on this, what would be an ideal window?

— notice that ideally, we want $E\{R_x^*(f)\} = R_x(f)$

— That's possible if $G(f) = \delta(f)$ ← Delta function

But $\delta(f) \overset{F.T}{\longleftrightarrow} 1$



But this requires infinite data in time domain.
so the ideal window is not possible, but we at least know that we would like our windows to be as close to the ideal window ($\delta(f)$) as possible.

(5) Periodogram Performance

(i) Bias $E\{R_x^*(f)\} = E\left\{\sum_{\tau=-n+1}^{n-1} r_x^*(\tau) e^{-i2\pi f\tau}\right\}$

$$= \sum \underbrace{E\{r_x^*(\tau)\}}\, e^{-i2\pi f\tau}$$

from chapter (3) → $$= \sum_{\tau=-n+1}^{n-1} \left(1 - \frac{|\tau|}{n}\right) r_x(\tau) e^{-i2\pi f\tau}$$

$$\neq R_x(f) = \sum_{\tau=-\infty}^{\infty} r_x(\tau) e^{-i2\pi f\tau}$$

so, the periodogram is biased!

(ii) $$\lim_{n\to\infty} E\{R_x^*(f)\} = \sum_{\tau=-\infty}^{\infty} r_x(\tau) e^{-i2\pi f\tau} = R_x(f)$$

so, its asymptotically unbiased

→ what this means is that you should collect a large amount of data.

(iii) As shown in sec 8.2.3

$$V\{R_x^*(f)\} \simeq \begin{cases} R_x^2(f) & 0 < |f| < 1/2 \\ 2R_x^2(f) & f = 0, \pm 1/2 \end{cases}$$

→ Thing to notice is that the variance does not decrease with n. So, periodo is always inconsistent and its variance remains high.

(iv) & (v) we need to look at the properties of the window.

— As shown in the book, for the periodogram (actually, an FD form of (i))

$$E\{R_x^*(f)\} = \int_{-1/2}^{1/2} R_x(u)\, K_n(f-u)\, du$$

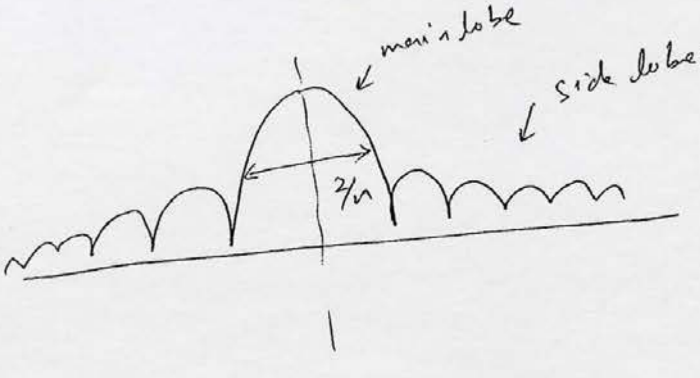
where $K_n(f) = \sum_{\tau=-n+1}^{n+1} k_n(\tau)\, e^{-i2\pi f \tau}$ ← Fejer's kernel (FD)

$k_n(\tau) = \max\left(0, 1 - \frac{|\tau|}{n}\right)$ ← ~~rectangular~~ triangular lag window (TD)

now $K_n(f) = \frac{\sin^2(\pi n f)}{n \sin^2(\pi f)} = n \cdot \left[\underbrace{\frac{\sin(n\pi f)}{n \sin(\pi f)}}_{D_n(f)} \right]^2$

$D_n(f)$ ← Dirichlet kernel

notice that $K_n(f)$ is a squared variant of the sinc function



(Paint brush analogy)

— Recall that the ideal window is a delta function, which means we want the mainlobe to be very narrow (sharp) and the sidelobes to be very low (no leakage).

— The main lobe affects resolution & side lobes affect masking.

— In Modified Periodograms we try to choose windows that have lower sidelobes (leading to less Bias) [but we loose resolution]